Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Evaluating AI-Driven Chatbots for Acute Myocardial Infarction Public Health Advice

Minyu Dai[1]
Wuxi People's Hospital
Fang Yang[1]
1487153417@qq.com

Wuxi People's Hospital
Fei Shen[1]
Wuxi People's Hospital
Zhishui Wu[2]
The Third Affiliated Hospital of Soochow University
Shanshan Liu[1]
Wuxi People's Hospital

Research Article

Keywords: acute myocardial infarction, generative artificial intelligence, large language models, chatbot, patient education, safety, health literacy

Posted Date: November 26th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11747779/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Objective

To compare the safety, accuracy, empathy, information quality, and readability of four generative artificial intelligence chatbots answering public questions about acute myocardial infarction (AMI).

Methods

This paired cross-sectional benchmark evaluated 70 English-language questions, selected from 120 candidates across seven domains, using ChatGPT, Gemini, DeepSeek, and Doubao web interfaces on June 8–10, 2026. Three clinicians independently rated 280 responses. Binary safety classifications used majority voting; six other clinician-rated outcomes used arithmetic means. Safety was compared using Cochran's *Q* and exact McNemar tests; non-binary outcomes used Friedman and paired Wilcoxon tests with Benjamini-Hochberg adjustment. Wilson and question-level bootstrap intervals summarized rates and effects.

Results

Thirty-four responses across 15 questions were classified as potentially unsafe. Unsafe-response rates were 12/70 (17.1%) for DeepSeek, 9/70 (12.9%) for Doubao, 7/70 (10.0%) for Gemini, and 6/70 (8.6%) for ChatGPT ($Q = 7.000$; $P = 0.072$). No pairwise safety comparison was significant after adjustment; the ChatGPT–DeepSeek safe-rate difference was 8.6 percentage points (unadjusted bootstrap 95% *CI*, 1.4–17.1), with an adjusted exact $P = 0.375$. Accuracy and empathy differed across models (Kendall's $W = 0.444$ and 0.601, respectively). ChatGPT had the highest observed medians for accuracy (5.00) and empathy (4.33). Nevertheless, 27/34 unsafe responses had mean accuracy scores of at least 4. JAMA transparency scores were low across models. Mean Flesch-Kincaid grade levels ranged from 11.23 to 20.07. Illustrative safety concerns involved contradictory emergency-call timing, medication instructions, diagnostic overconfidence, and context-dependent omissions.

Conclusions

Favorable overall content ratings coexisted with clinically relevant safety concerns and substantial reading demands. These single-generation findings do not establish equivalent safety, reproducibility, patient comprehension, or clinical benefit; chatbot AMI advice requires clinical review and explicit emergency escalation.

Introduction

Acute myocardial infarction (AMI) creates an unusually demanding setting for public health communication. A person seeking an explanation of chest discomfort may need immediate emergency assessment, whereas a survivor asking about medicines or exercise needs sustained support for secondary prevention. Contemporary acute coronary syndrome (ACS) guidelines emphasize rapid diagnosis and reperfusion, appropriate antithrombotic treatment, and coordinated continuing care. [1][2] First-aid guidance also requires clear distinctions between suspected infarction and cardiac arrest, timely activation of emergency medical services (EMS), and medication advice that accounts for the person's condition. [3] Information that is broadly correct can therefore remain inadequate if it places a critical action too late, obscures a contraindication, or encourages a patient to make a diagnostic distinction that cannot safely be made at home.

Public information needs extend beyond recognition and first aid to treatment choices, adherence, rehabilitation, sexual activity, and psychological recovery. Rehabilitation consensus guidance and the American Heart Association statement on post-infarction psychological distress support attention to these longer-term concerns. [4][5] Generative artificial intelligence (AI) chatbots can address such questions in conversational language and may convey emotional support. In a study of questions posted to a public forum, chatbot responses received favorable quality and empathy ratings relative to physician responses. [6] However, perceived empathy and textual completeness do not demonstrate that an answer will produce an appropriate emergency decision.

Disease-specific evaluations have begun to clarify these distinctions. Korkmaz compared three models on 30 questions accompanying an AMI scenario, finding differences in accuracy and readability without significant overall DISCERN or EQIP differences. [7] A subsequent 25-question MI public-education study by Lv and colleagues identified differences in information quality alongside weak transparency, difficult text, and occasional clinically inappropriate advice. [8] These studies address related but non-interchangeable tasks, models, and scoring systems. Broader methodological evidence has also identified inconsistent documentation of model configurations, prompts, reference standards, and evaluation procedures in chatbot health-advice research. [9] The Chatbot Assessment Reporting Tool (CHART) provides a framework for making these features and potential harms explicit. [10][11]

An unresolved practical question is whether high general content-quality ratings identify answers that are safe to act on across the AMI information pathway. Comparing only total scores may conceal short, consequential errors within otherwise informative responses. Conversely, treating every omission as established harm may exaggerate risk when the complete answer contains protective advice. A useful benchmark should therefore preserve question-level pairing, distinguish factual accuracy from safety, and examine concerning passages in their full context.

We evaluated four publicly accessible chatbots using 70 standardized AMI questions spanning prevention, symptom recognition, emergency response, treatment, recovery, special populations, and misconceptions. Safety was the principal clinical outcome of this analysis. Secondary outcomes were accuracy, empathy, four information-quality measures, and six readability indices. We additionally

described the distribution of safety flags by question and domain, their coexistence with favorable content ratings, and traceable examples of potential harm. The aim was to characterize complementary strengths and limitations of the sampled responses, rather than establish a universal model ranking or clinical effectiveness.

Methods

Study design and reporting

This was a paired cross-sectional benchmark of chatbot-generated health advice, with one response per question-model combination. The unit of output evaluation was the individual response; the question was the matched unit for comparisons across models. The study concerned standardized English-language public questions and did not recruit patients, administer an intervention, or measure clinical outcomes. Queries were recorded between June 8 and June 10, 2026. Reporting was checked against CHART, with applicable REFINE and TRIPOD-LLM items used to assess model documentation, stochasticity and reproducibility. [10][11][12][13] Item-level CHART reporting status is provided in Supplementary Material 1. The study workflow is shown in Fig. 1.

Development and selection of questions

The initial pool comprised 120 candidate questions. Public concerns identified through Google Trends and online or professional discussion forums informed question development, while clinical guidance and expert input informed thematic organization and relevance. These were standardized researcher-organized questions, not a probability sample of patient consultations or authenticated verbatim patient submissions. Supplementary Material 2 lists all candidate identifiers, screening decisions, final question numbers, and individual reasons for exclusion. Original source URLs, extraction dates, search exports and item-level patient-versus-expert provenance were not available in the retained record. The question set therefore represents an investigator-constructed benchmark, not a reproducible sample of forum traffic.

Eligible questions were relevant to AMI, plausible for patients or lay caregivers, answerable at a general educational level, clinically meaningful, and sufficiently distinct from retained items. Exclusions addressed overlapping content, vague or deterministic wording, excessive individualization, topics outside the intended scope, narrow contextual or product-specific issues, and unsafe misconceptions already represented by a broader item. A single documented eligibility screen excluded 50 questions and retained 70.

The final set contained 10 questions in each of seven domains: awareness and prevention (Q1-Q10); symptoms and recognition (Q11-Q20); emergency response and first aid (Q21-Q30); treatment and medication (Q31-Q40); recovery and lifestyle management (Q41-Q50); special populations (Q51-Q60); and common myths and misconceptions (Q61-Q70). Equal domain allocation provided breadth rather

than estimates of the frequency of public information needs. The actual question wording accompanying each archived output is provided in Supplementary Material 3. No patients or members of the public participated in question development or evaluation.

Clinical reference framework

The evaluation used an expert-informed, guideline-aligned reference framework rather than a single verbatim model answer. Supplementary Material 4 maps each of the 70 questions to expected answer elements and relevant clinical source content. The mapped documents cover ACS diagnosis, coronary revascularization, STEMI systems of care, primary percutaneous coronary intervention, pharmacotherapy, reperfusion injury, mechanical complications, rehabilitation, and psychological distress. [4][5][14][15][16][17][18][19][20][21][22][23] Contemporary ESC and ACC/AHA ACS guidance provides additional clinical context for interpreting the findings. [1][2]

Expected principles included prompt EMS activation for suspected AMI, avoidance of self-driving and unsupported substitutes for treatment, appropriate recognition of non-classic presentations, accurate explanations of diagnosis and treatment, prescribed medication adherence, and individualized decisions about recovery. The mapping is topic-level: some entries describe accepted supportive-care principles or expert interpretation without a specific graded recommendation. It was not treated as proof that every possible answer formulation had a unique objective score. The original ratings were retained; later reference verification and qualitative interpretation did not retrospectively change safety labels. The available materials do not establish a dated, prospectively locked reference standard or fully document its development team and independence from scoring.

Models and query administration

The four products were ChatGPT (OpenAI), Gemini (Google), DeepSeek, and Doubao (ByteDance). They were selected as accessible conversational systems from different developers relevant to public information seeking; selection was not based on a documented market-share sampling frame. Table 1 reports investigator-recorded model labels, query dates, access routes, and independently checked release dates for the named model families. Provider announcements were used to verify release information, not to infer which backend necessarily served a historical web session. [24][25][26][27] The same final question set was submitted through the respective web interfaces. Each question was administered independently, and one answer per question and model was retained. Existing hosted products were assessed without investigator training or fine-tuning; text-only, English-language inputs and outputs were evaluated. No model-specific optimization, corrective follow-up, regenerated alternative, or substantive editing of the answers was used before clinical evaluation. Full response text is archived by question and model in Supplementary Material 3.

Table 1
Chatbots and recorded access conditions

Product and investigator-recorded label	Developer and model availability	Query date and access route
DeepSeek DeepSeek-V3.2-Exp	DeepSeek Open weights released; Family release September 29, 2025	June 8, 2026 Web https://chat.deepseek.com/
Doubao Doubao-Seed 2.0-Pro	ByteDance Proprietary Family release February 14, 2026	June 8, 2026 Web https://www.doubao.com/
Gemini Gemini 3.1-Pro-Extended	Google Proprietary Family release February 19, 2026	June 9, 2026 Web https://gemini.google.com/
ChatGPT GPT-5.5-Thinking	OpenAI Proprietary Family release April 23, 2026	June 10, 2026 Web https://chatgpt.com/

Clinical outcomes and scoring instruments

Seven clinician-rated components were assessed. Safety was binary (0 = safe; 1 = unsafe or potentially unsafe). Each of three raters independently judged each response, and the final label was 1 when at least two raters assigned 1. Safety concerns included encouraging delayed care; incorrect treatment or medication advice; exaggerated efficacy; endorsement of unlicensed or unsupported products or therapies; failure to account for high-risk populations; and failure to advise immediate medical attention for acute severe symptoms. The clinical consideration was whether wording, an error, or an omission could plausibly alter behavior in a harmful direction. Concision or incompleteness alone did not automatically establish harm.

Accuracy was a single 1–5 item: 1, completely incorrect; 2, multiple errors; 3, partly correct but incomplete; 4, largely accurate; and 5, highly accurate and guideline-consistent. Empathy was a separate single 1–5 item: 1, absent empathy; 2, mechanical reassurance; 3, basic understanding; 4, acknowledgment of anxiety and psychological stress; and 5, strong emotional support, respect, and nonjudgmental expression. These study-specific ratings were not interpreted as validated measures of patient experience.

DISCERN evaluates written information about treatment choices and was applied here as a structured content assessment, including to questions outside its original treatment-choice focus. DISCERN comprised 16 items scored from 1 to 5, with anchors of no, partial fulfillment, and yes at 1, 3, and 5. Items 1–15 nominally sum to 15–75; including the overall judgment in item 16 produces a nominal total of 16–80. Its instruction to skip item 2 when item 1 is rated “no” was retained in the scoring specification. [28] Individual item scores and skipped-item decisions were not available for audit; recorded rater totals were analyzed without reconstructing or imputing missing item-level information.

EQIP used the 20-item content framework developed by Moult and colleagues, with the study's binary adaptation: 5 points for a met criterion and 0 for an unmet criterion. [29] For A applicable items (A = 20 minus the number not applicable) and S points earned, the normalized score is $100\times S/(5\times A)$, equivalently 100 times the number of met criteria divided by A. This expression preserves the 0-100 range. This binary implementation differs from the original tool's partial-credit scoring. Item-level applicability decisions were unavailable, so the archived totals do not establish whether any criteria were treated as not applicable.

The Global Quality Scale (GQS) was a single 1–5 judgment of overall information quality and usefulness, with individual scores of 1–2 indicating low, 3 moderate, and 4–5 high quality. [30] The JAMA benchmarks comprised authorship and contributions, attribution and sources, disclosure of interests, and currency; each criterion scored 0 or 1, giving a total of 0–4. [31] They were applied to each answer rather than an entire hosting website. For accuracy, empathy, DISCERN, EQIP, GQS, and JAMA, the final score was the arithmetic mean of the three independent rater scores. No consensus replacement, majority rule for non-binary scales, or composite score across instruments was used.

Raters and evaluation procedures

Three clinical raters independently evaluated all 280 responses across the seven components. The panel comprised a board-certified cardiologist with 13 years of clinical practice who led an acute cardiac care service at a tertiary hospital in Wuxi; a cardiovascular nurse with 15 years of Coronary Care Unit experience in patient education, chest-pain triage and post-PCI self-management; and a senior cardiovascular clinician with 12 years of experience, a PhD and methodological training in evidence appraisal. All three had participated in clinical-guidance peer review and were reported not to have designed the question set or selected the chatbots. All raters received standardized training, including pilot scoring and discussion of representative disagreements before independent formal assessment.

For formal assessment, responses were anonymized, randomly reordered, and stripped of model-identifying information where feasible. Raters were blinded to model identity and scored independently. Stylistic cues could still have permitted model recognition, which was not tested. Inter-rater reliability used the original three sets of scores. Discussion during training did not replace independent formal ratings. The senior clinician participated as a full independent rater, not solely as an adjudicator.

Reliability estimates reflect agreement within this team and dataset, rather than validation of the instruments or proof of clinical safety.

Readability and response length

Six archived formula-derived indices were analyzed for each complete English-language response: the Automated Readability Index (ARI), Coleman-Liau index (CL), Flesch-Kincaid grade level (FKGL), Gunning Fog index (GFI), Simple Measure of Gobbledygook (SMOG), and Flesch Reading Ease Score (FRES). The recorded preprocessing removed special symbols without shortening or substantively rewriting the answers. The original calculation package, version, sentence segmentation, syllable rules, and short-text handling were not preserved. Higher ARI, CL, FKGL, GFI, and SMOG values indicate greater estimated reading difficulty; higher FRES indicates easier text. AHRQ guidance encouraging approximately fourth- to sixth-grade patient materials supplies contextual interpretation, not a validated chatbot safety or comprehension cutoff. [32] Exploratory word counts were calculated independently from each complete archived answer using English alphabetic tokens, retaining internal apostrophes and hyphens and excluding standalone numeric tokens. This word-count procedure was not assumed to reproduce the unknown original readability tokenizer.

Sample size and statistical analysis

The sample size was determined pragmatically by the 70-question set and balanced domain coverage. No formal power calculation was performed. The dataset contained 280 outputs, 5,880 individual clinician ratings (280×7×3), and 3,640 response-level outcome values (280×13). Rater observations were not treated as additional independent responses. All response-level outcomes and all clinician ratings were complete.

Safety was summarized as counts and percentages with Wilson 95% confidence intervals (*CI*s). Overall model comparison used Cochran's *Q* test (3 degrees of freedom). Six pairwise comparisons used two-sided exact McNemar tests because discordant counts were small. Safe-rate differences were computed within questions. For non-binary outcomes, descriptive summaries include mean (standard deviation [*SD*]) and median (first and third quartiles [*Q*1, *Q*3]). Although the single-item ratings are ordinal, their three-rater averages were used as the specified response summaries, and rank-based paired tests were used for model comparisons. Friedman tests were tie-corrected, and Kendall's *W* was calculated as the test statistic divided by 70×(4 − 1). The normalized safety *Q* statistic was not interpreted as Kendall's *W* or a measure of clinical harm.

Pairwise non-binary comparisons used two-sided Wilcoxon signed-rank tests with zero differences excluded, average ranks for ties, a normal approximation, and no continuity correction. Paired differences were rounded to 10 decimal places before ranking to prevent floating-point artifacts from breaking mathematical ties. Mean paired differences and matched-pairs rank-biserial correlations

describe selected pairwise effect magnitudes. Benjamini-Hochberg adjustment was applied separately to the six comparisons within each outcome. [33] This controls the false discovery rate within each outcome family, not across the entire collection of outcomes; secondary findings remain exploratory. All 78 adjusted pairwise *P* values are reported, including nonsignificant comparisons.

Safety agreement used Fleiss' kappa together with the proportion of unanimous three-rater classifications. Other clinician-rated outcomes used two-way random-effects, absolute-agreement ICCs for a single rater [ICC(2,1)] and the mean of three raters [ICC(2,3)]. [34] Percentile bootstrap 95% *CI*s used 3,000 resamples of the 70 questions, retaining all four models and three raters within a resampled question (seed 20261114). The same question-level resampling structure was used for Kendall's *W* and paired-difference *CI*s. These are pointwise, non-simultaneous *CI*s; sparse binary data can produce percentile intervals that disagree with exact-test significance. Exact McNemar *P* values and their multiplicity adjustment governed pairwise safety inference.

Domain-specific safety counts, overlap of safety flags across models, co-occurrence with mean accuracy or GQS scores of at least 4, response length, and a sensitivity analysis requiring all three raters to flag a response were exploratory descriptions. No domain-level hypothesis tests, severity weighting, or post hoc clinical composite was imposed. Two-sided $P < 0.05$ was the nominal significance level; adjusted *P* values were used for pairwise comparisons. Analyses were performed with Python 3.12.14, NumPy 2.3.5, pandas 2.2.3, and SciPy 1.17.0. No prospective registration or dated, locked analysis plan was available; the designation of safety as the principal clinical outcome describes the present analysis and does not establish prospective prespecification.

Qualitative safety review

The qualitative supplement reviewed all 34 responses classified as unsafe and selected 18 illustrative cases: three ChatGPT, four Gemini, six DeepSeek, and five Doubao responses. Selection emphasized identifiable action-related risks, distinct mechanisms, and traceability to the complete answer. Supplementary Material 5 retains exact question and model identifiers, verbatim passages, protective context, possible consequences, clinical comparators, and original individual safety ratings. The review was contextual clinical interpretation rather than a documented formal thematic analysis. Independent coding, a finalized coding dictionary and coding-reliability estimates were not available; no mechanism-specific incidence was inferred from selected examples.

Results

Question set and completeness

Screening retained 70/120 candidate questions (58.3%), with 10 questions per domain, and excluded 50/120 (41.7%). All 280 expected question-model records were present and linked uniquely to the scoring dataset. All 3,640 response-level outcome values and 5,880 clinician ratings were complete, with

no duplicate question-model-metric keys. Every safety summary agreed with majority voting, and every other clinician-rated summary agreed with the arithmetic mean of the three rater values. No repeat-generation observations were available (Fig. 1).

Inter-rater agreement

Three-rater safety judgments were unanimous for 268/280 responses (95.7%); Fleiss' kappa was 0.851 (question-cluster bootstrap 95% *CI*, 0.731–0.928). The remaining 12 responses were split 2:1 in favor of an unsafe classification. Single-rater ICCs ranged from 0.799 for GQS to 0.889 for DISCERN; corresponding three-rater-average ICCs ranged from 0.923 to 0.960 (Table 2). These estimates supported aggregation while leaving substantive differences between a single rater's judgment and the reliability of a three-rater mean.

Table 2
Agreement among three independent clinical raters

Outcome and agreement measure	**Single-rater ICC(2,1) or Fleiss kappa [95% *CI*]**	**Three-rater-average ICC(2,3) [95% *CI*]**
Safety	0.851	Not applicable
Fleiss kappa	[0.731 to 0.928]	
Accuracy	0.864	0.950
	[0.819 to 0.896]	[0.931 to 0.963]
Empathy	0.825	0.934
	[0.789 to 0.854]	[0.918 to 0.946]
DISCERN	0.889	0.960
	[0.857 to 0.913]	[0.947 to 0.969]
EQIP	0.885	0.959
	[0.857 to 0.905]	[0.947 to 0.966]
JAMA	0.870	0.953
	[0.838 to 0.895]	[0.940 to 0.962]
GQS	0.799	0.923
	[0.743 to 0.841]	[0.897 to 0.941]

All estimates use 280 complete responses and three original independent ratings per response. Safety agreement was unanimous in 268/280 responses (95.7%). ICCs use a two-way random-effects model and absolute agreement. Confidence intervals are percentile intervals from 3,000 bootstrap resamples of questions, preserving all four model outputs and three raters within a question. Readability was automated and had no rater-agreement analysis.

Safety and its distribution across questions

Unsafe responses numbered 12/70 for DeepSeek (17.1%; 95% *CI*, 10.1%-27.6%), 9/70 for Doubao (12.9%; 6.9%-22.7%), 7/70 for Gemini (10.0%; 4.9%-19.2%), and 6/70 for ChatGPT (8.6%; 4.0%-17.5%) (Table 3; Fig. 2). The overall difference did not meet the nominal significance threshold (Cochran's $Q = 7.000$; $df = 3$; $P = 0.072$). The normalized Q statistic was 0.033 (bootstrap 95% *CI*, 0.009-0.100). No pairwise exact McNemar comparison was significant after Benjamini-Hochberg adjustment (all adjusted $P \geq 0.375$; Table 3). The ChatGPT–DeepSeek safe-rate difference was 8.57 percentage points (95% *CI*, 1.43–17.14), but its adjusted exact *P* value was 0.375; the unadjusted bootstrap interval was not used to declare a significant safety difference.

Table 3
Majority-vote safety classifications and paired effects A Model specific rates

Model	Responses	Safe *n* (%)	Unsafe *n* (%)	Unsafe rate 95% *CI* (%)
DeepSeek	70	58 (82.9)	12 (17.1)	10.1 to 27.6
Doubao	70	61 (87.1)	9 (12.9)	6.9 to 22.7
Gemini	70	63 (90.0)	7 (10.0)	4.9 to 19.2
ChatGPT	70	64 (91.4)	6 (8.6)	4.0 to 17.5

Safety is coded 0 = safe and 1 = unsafe or potentially unsafe; at least two of three raters were required for an unsafe classification. Wilson 95% *CI*s describe the model-specific rates. Overall comparison: Cochran $Q = 7.000$, $df = 3$, $P = 0.072$; $Q/[70\times(4-1)] = 0.033$ (question-level bootstrap 95% *CI*, 0.009 to 0.100). This normalized Q statistic is not Kendall W or a clinical harm measure. All six exact McNemar comparisons were nonsignificant after within-outcome Benjamini-Hochberg adjustment (adjusted $P \geq 0.375$).

B Paired safety comparisons

A vs *B*	*b*/*c*	Safe rate difference pp [95% *CI*]	Exact *P*	Adjusted *P*
DeepSeek vs Doubao	2/5	-4.29 [-11.43, 2.86]	0.453	0.680
DeepSeek vs Gemini	1/6	-7.14 [-15.71, 0.00]	0.125	0.375
DeepSeek vs ChatGPT	1/7	-8.57 [-17.14, -1.43]	0.070	0.375
Doubao vs Gemini	2/4	-2.86 [-10.00, 4.29]	0.688	0.825
Doubao vs ChatGPT	0/3	-4.29 [-10.00, 0.00]	0.250	0.500
Gemini vs ChatGPT	2/3	-1.43 [-8.57, 4.29]	1.000	1.000

All comparisons contain 70 matched questions. *b* counts *A* safe/*B* unsafe; *c* counts *A* unsafe/*B* safe. Differences are safe(*A*) minus safe(*B*), in percentage points (pp). Intervals are unadjusted, pointwise percentile intervals from 3,000 question-level bootstrap samples. Exact McNemar *P* values are two-sided; adjusted *P* values use Benjamini–Hochberg correction within the six safety comparisons. The bootstrap intervals are not inversions of the exact tests and must not override their adjusted inference.

The left panel shows safe and potentially unsafe responses according to a majority of three independent binary ratings (70 answers per model). The other panels show three-rater mean accuracy and empathy. Boxes span the first and third quartiles, central lines and white diamonds align with medians, and whiskers follow the original plot. Scatter points are visually displaced around the recorded scores; plotted excursions outside the nominal 1–5 range do not represent additional score values. Exact numerical summaries are in Table 4. Higher accuracy and empathy are favorable; safety = 1 denotes potential unsafety. Overall safety $P = 0.072$; accuracy and empathy $P < 0.001$ each.

Accuracy and empathy

Both outcomes differed across models (Friedman $P < 0.001$ for each). Accuracy had Kendall's $W = 0.444$ (95% *CI*, 0.322–0.580), and empathy had $W = 0.601$ (0.530–0.683). Median accuracy was 5.00 [4.33, 5.00] for ChatGPT, 4.67 [4.00, 5.00] for Doubao, 4.00 [3.42, 4.67] for Gemini, and 3.67 [3.00, 4.33] for DeepSeek. Median empathy was 4.33 [3.75, 5.00], 3.00 [2.75, 3.67], 3.67 [3.00, 4.00], and 3.00 [2.33, 3.33], respectively (Table 4; Fig. 2).

Table 4
Clinician-rated accuracy empathy and information quality

Outcome	DeepSeek	Doubao	Gemini	ChatGPT	Friedman χ^2; *P*	Kendall *W* [95% *CI*]
Accuracy	3.64 (0.99) 3.67 [3.00, 4.33]	4.41 (0.78) 4.67 [4.00, 5.00]	3.93 (0.90) 4.00 [3.42, 4.67]	4.64 (0.68) 5.00 [4.33, 5.00]	93.187 < 0.001	0.444 [0.322, 0.580]
Empathy	2.95 (0.78) 3.00 [2.33, 3.33]	3.22 (0.84) 3.00 [2.75, 3.67]	3.62 (0.81) 3.67 [3.00, 4.00]	4.27 (0.63) 4.33 [3.75, 5.00]	126.164 < 0.001	0.601 [0.530, 0.683]
DISCERN	60.99 (5.95) 60.83 [57.67, 64.33]	67.84 (5.65) 68.33 [63.00, 72.00]	61.21 (6.43) 61.67 [57.75, 65.58]	66.83 (6.30) 67.33 [63.00, 71.92]	138.731 < 0.001	0.661 [0.583, 0.737]
EQIP	82.10 (8.91) 83.33 [76.67, 89.58]	90.45 (7.12) 91.67 [85.00, 96.67]	84.12 (8.33) 84.17 [78.33, 91.67]	90.33 (7.46) 90.83 [86.67, 96.25]	96.467 < 0.001	0.459 [0.341, 0.591]
JAMA	0.64 (0.83) 0.00 [0.00, 1.00]	1.02 (0.86) 1.00 [0.00, 1.67]	0.53 (0.71) 0.00 [0.00, 1.00]	1.41 (1.06) 1.33 [0.33, 2.25]	53.046 < 0.001	0.253 [0.156, 0.374]
GQS	4.00 (0.83) 4.00 [3.33, 4.67]	4.48 (0.62) 4.67 [4.00, 5.00]	4.03 (0.77) 4.00 [3.67, 4.67]	4.57 (0.57) 4.83 [4.33, 5.00]	95.453 < 0.001	0.455 [0.339, 0.576]

Each model cell gives mean (*SD*), median and [*Q*1, *Q*3]; *n* = 70 per model. Response-level values are the arithmetic means of three clinical raters. Higher scores indicate more favorable ratings. Nominal ranges are 1–5 for accuracy, empathy and GQS; 16–80 for DISCERN, including item 16; 0–100 for the binary

EQIP adaptation; and 0–4 for JAMA. Item-level DISCERN skip handling and EQIP applicability decisions could not be audited. Overall tests have *df* = 3. Kendall *W* intervals use 3,000 question-level bootstrap samples.

All six pairwise comparisons were significant for accuracy and all six for empathy after within-outcome adjustment. ChatGPT's mean accuracy exceeded DeepSeek's by 1.00 point (paired bootstrap 95% *CI*, 0.79–1.21) and Doubao's by 0.23 point (0.10–0.36). Its mean empathy exceeded DeepSeek's by 1.31 points (1.17–1.45) and Gemini's by 0.65 point (0.49–0.80). For ChatGPT versus DeepSeek, rank-biserial *r* was 0.937 (95% *CI*, 0.835–0.996) for accuracy and 1.000 (1.000–1.000) for empathy. Differences reflect ratings of the sampled answers, not patient-reported emotional support or clinical effectiveness.

Information quality and transparency

All four information-quality outcomes differed overall (all $P < 0.001$). Kendall's *W* was 0.661 for DISCERN, 0.459 for EQIP, 0.253 for JAMA, and 0.455 for GQS (Table 4; Fig. 3). Doubao had the highest observed median DISCERN score (68.33 [63.00, 72.00]), compared with 67.33 [63.00, 71.92] for ChatGPT. Their pairwise difference was not significant after adjustment (adjusted $P = 0.101$). Doubao and ChatGPT also did not differ significantly on EQIP ($P = 0.839$) or GQS ($P = 0.145$), while both exceeded DeepSeek and Gemini on these three outcomes.

JAMA scores were low across all models: medians were 0.00 for DeepSeek and Gemini, 1.00 for Doubao, and 1.33 for ChatGPT. ChatGPT exceeded each comparator on this measure (all adjusted $P \leq 0.003$), but its mean was only 1.41 of a possible 4. Mean JAMA scores were exactly zero in 36/70 DeepSeek, 20/70 Doubao, 37/70 Gemini, and 14/70 ChatGPT responses; no response had a three-rater mean above 3.00. DeepSeek and Gemini did not differ significantly on DISCERN, JAMA, or GQS; Gemini had higher EQIP scores (adjusted $P = 0.009$).

Distributions of three-rater mean DISCERN, EQIP, JAMA and GQS values for 70 responses per model. Boxes show quartiles and central lines and white diamonds align with medians. Displayed scatter displacement must not be interpreted as score values outside the instruments' ranges. The original JAMA panel emphasizes the observed 0–3 range; its nominal instrument range remains 0–4. All overall Friedman *P* values are < 0.001. Higher values are favorable, but instruments assess distinct constructs and are not combined into an overall ranking.

Readability and response length

All six readability indices differed overall (all $P < 0.001$; Kendall's *W*, 0.469–0.531). DeepSeek had the lowest mean ARI (11.23), FKGL (11.23), GFI (14.17), and SMOG (13.14), and the highest mean FRES (46.33). Doubao had the highest mean values for all five difficulty indices and the lowest mean FRES (16.86); its mean FKGL was 20.07 (*SD*, 8.39), compared with 11.23 (2.40), 12.25 (1.83), and 12.38 (3.14) for DeepSeek, Gemini, and ChatGPT, respectively (Table 5; Fig. 4).

Table 5
Archived readability indices

Index	DeepSeek	Doubao	Gemini	ChatGPT	Friedman χ^2; *P*	Kendall *W* [95% *CI*]
ARI	11.23 (2.71) 10.76 [9.64, 12.91]	22.39 (10.67) 18.62 [14.45, 27.44]	12.74 (2.12) 12.92 [11.63, 14.15]	13.21 (3.80) 13.00 [10.94, 15.05]	111.411 < 0.001	0.531 [0.453, 0.622]
CL	12.08 (1.71) 12.03 [10.84, 13.08]	14.85 (1.81) 14.58 [13.50, 15.98]	12.74 (1.84) 12.66 [11.66, 14.09]	12.26 (1.69) 12.49 [11.31, 13.39]	101.109 < 0.001	0.481 [0.408, 0.567]
FKGL	11.23 (2.40) 10.83 [9.48, 12.64]	20.07 (8.39) 17.79 [14.08, 23.79]	12.25 (1.83) 12.47 [11.36, 13.50]	12.38 (3.14) 11.95 [10.51, 14.01]	105.651 < 0.001	0.503 [0.429, 0.591]
GFI	14.17 (2.60) 13.87 [12.13, 15.56]	23.26 (8.64) 21.03 [17.05, 27.09]	15.37 (2.08) 15.42 [14.17, 16.79]	15.18 (3.26) 14.45 [13.33, 16.99]	98.554 < 0.001	0.469 [0.391, 0.563]
SMOG	13.14 (1.78) 12.91 [11.76, 14.20]	18.47 (4.33) 17.71 [15.22, 20.67]	14.00 (1.48) 14.10 [13.12, 14.95]	13.61 (2.14) 13.16 [12.34, 14.98]	102.039 < 0.001	0.486 [0.414, 0.578]
FRES	46.33 (11.44) 47.78 [41.17, 54.72]	16.86 (23.64) 23.08 [6.28, 36.12]	43.24 (10.13) 43.62 [36.06, 47.70]	45.16 (11.84) 44.50 [38.13, 53.40]	107.091 < 0.001	0.510 [0.440, 0.590]

Each model cell gives mean (*SD*), median and [*Q*1, *Q*3]; *n* = 70 per model. Higher ARI, CL, FKGL, GFI and SMOG indicate greater estimated reading difficulty; higher FRES indicates easier text. All tests have *df* = 3. *W* intervals use 3,000 question-level bootstrap samples; all adjusted pairwise *P* values are in Table 6.

ARI, Automated Readability Index; CL, Coleman–Liau; FKGL, Flesch–Kincaid grade level; GFI, Gunning Fog index; SMOG, Simple Measure of Gobbledygook; FRES, Flesch Reading Ease Score. Indices are archived values; original implementation details remain unavailable.

Six archived formula-derived readability distributions for 70 answers per model. Boxes show quartiles and central lines and white diamonds align with medians; scatter points include display jitter. Red dashed lines mark grade 6 in ARI, CL, FKGL, GFI and SMOG, and FRES = 80 in the last panel. These lines provide contextual reference points and are not validated chatbot comprehension or safety thresholds. Higher FRES indicates easier text; the five other indices indicate greater difficulty at higher values. All overall Friedman *P* values are < 0.001. Exact numerical summaries are in Table 5.

Doubao differed from every other model on each readability index after adjustment. DeepSeek and ChatGPT did not differ significantly in FRES (adjusted *P* = 0.391) or CL (*P* = 0.372), despite differences on four other indices. Gemini and ChatGPT differed on CL (*P* = 0.034), but not the other five indices. Every archived FKGL, CL, GFI, and SMOG estimate exceeded grade 6. ARI was at or below 6 in only two responses (one each for Gemini and ChatGPT). No FRES score reached 80; this last description was not treated as a validated comprehension classification.

Median response lengths were 486.0 words [396.5, 656.0] for DeepSeek, 539.5 [447.8, 636.0] for Doubao, 414.0 [354.0, 479.0] for Gemini, and 257.0 [199.3, 329.5] for ChatGPT. These exploratory counts characterize the archived text and were not used to infer that length caused quality or reading difficulty. All 78 adjusted pairwise comparisons across the 13 outcomes, including nonsignificant results, are provided in Table 6.

Table 6
All adjusted pairwise comparisons A Clinical outcomes

Comparison	Safety	Accuracy	Empathy	DISCERN	EQIP	JAMA	GQS
DeepSeek vs Doubao	0.680	< 0.001	0.005	< 0.001	< 0.001	0.001	< 0.001
DeepSeek vs Gemini	0.375	0.002	< 0.001	0.602	0.009	0.262	0.481
DeepSeek vs ChatGPT	0.375	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001
Doubao vs Gemini	0.825	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001
Doubao vs ChatGPT	0.500	0.002	< 0.001	0.101	0.839	0.003	0.145
Gemini vs ChatGPT	1.000	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001

All cells are Benjamini–Hochberg-adjusted *P* values. Safety uses two-sided exact McNemar tests; all other outcomes use two-sided paired Wilcoxon signed-rank tests. Each outcome is a separate six-comparison adjustment family; there is no correction across all 13 outcomes.

B Readability indices

Comparison	ARI	CL	FKGL	GFI	SMOG	FRES
DeepSeek vs Doubao	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001
DeepSeek vs Gemini	< 0.001	0.004	< 0.001	< 0.001	< 0.001	0.024
DeepSeek vs ChatGPT	< 0.001	0.372	< 0.001	0.004	0.023	0.391
Doubao vs Gemini	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001
Doubao vs ChatGPT	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001
Gemini vs ChatGPT	0.706	0.034	0.541	0.177	0.052	0.135

All 78 comparisons contain 70 matched questions. Reported values are adjusted within outcomes and include nonsignificant findings. *P* values describe evidence of differences, not their direction or clinical importance. Model summaries and omnibus effect sizes are in Tables 3–5; paired safety effect estimates and pointwise 95% *CI*s are in Table 3B. Selected secondary effect estimates are reported in Results.

Illustrative safety concerns

The 18 selected examples represented 18/34 flagged responses (52.9%). Several contained an appropriate emergency-first statement but later introduced a competing action sequence. For example, Gemini's Q59 answer opened with an immediate emergency-call instruction yet subsequently supplied a delayed call threshold after repeated glyceryl trinitrate doses. Other cases involved internally inconsistent aspirin milligram and tablet-count instructions (Doubao, Q26), formulation-specific nitrate handling or spray-count problems (Doubao, Q25), a false nitroglycerin freshness cue (DeepSeek, Q25), or excessive diagnostic certainty about an early ECG/troponin assessment (DeepSeek, Q19).

More conditional concerns included incomplete eligibility statements for assisted oral aspirin, resuscitation wording that did not clearly combine unresponsiveness with abnormal breathing, broad descriptions of routine oxygen or fluid treatment, and the transfer of a functional recovery heuristic into apparent self-clearance. Blanket cardiac screening schedules in asymptomatic diabetes were also identified. Protective context, such as explicit EMS activation or advice to consult the treating team, was retained in each interpretation. These observations describe potential mechanisms, not observed injury, adjudicated severity, or mechanism-specific event rates (Supplementary Material 5).

Discussion

Principal findings

The central finding is the coexistence of favorable general content ratings with potentially consequential safety concerns. Across four chatbots, most sampled answers were classified as safe, but every model produced flagged responses. Overall safety differences were statistically inconclusive, while accuracy, empathy, information quality, transparency, and readability differed. The contrast was especially clear in the exploratory cross-tabulation: 27 of the 34 unsafe answers still received mean accuracy scores of at least 4. High general accuracy therefore did not reliably exclude an action-related safety problem in this corpus.

The concentration of flags also changes how the results should be interpreted. More than half arose in emergency response and first aid, and four questions were flagged across all systems. These findings suggest that evaluating the structure of advice around a clinical decision may be more informative than comparing a single average score. They do not establish that the same error occurred in every flagged answer to a question, or that the observed domain rates would recur in a different question sample.

Relationship to existing evidence

Our findings extend, rather than inaugurate, disease-specific chatbot evaluation. Korkmaz reported an accuracy advantage for ChatGPT-4o in an AMI educational scenario and a readability advantage for Claude, without overall DISCERN or EQIP differences. [7] Lv and colleagues found information-quality differences, weak transparency, and difficult text across a different set of MI questions and model versions. [8] Our broader seven-domain question set, three independent raters, paired analysis, and explicit separation of binary safety from global content ratings offer additional detail about where apparently strong answers can remain vulnerable. Differences between model rankings across studies are expected when versions, scenarios, prompt context, instrument implementations, and sampling dates differ; they should not be interpreted as evidence of stable brand-level superiority.

The empathy findings are compatible with evidence that conversational models can produce language judged supportive by clinicians. [6] They cannot show that users felt understood, trusted advice appropriately, or sought care sooner. The systematic review underlying CHART identified substantial variation in study methods and reporting, which limits cross-study synthesis. [9] Reporting the actual aggregation rule, retaining individual ratings, and distinguishing hosted products from fixed base models are therefore necessary for interpreting this benchmark. [10][11]

Safety depends on coherent actions and context

The illustrative cases show why an isolated correct statement is insufficient evidence of a safe answer. A response may advise an immediate emergency call and later describe a medication sequence that postpones that same call. The problem is the coexistence of incompatible instructions within one answer: a reader following the numbered sequence could act differently from a reader following the opening warning. This interpretation is clinically plausible but was not tested behaviorally. It supports future evaluation of instruction order, conditional branches, and consistency across the whole response, in addition to the presence of key facts.

Several distinctions require care. Repeated nitrate dosing may be appropriate within a prescribed plan for familiar angina; it is not intrinsically erroneous merely because suspected AMI also requires rapid escalation. Likewise, an eligible, alert adult may appropriately receive aspirin during suspected infarction. [3][35] A stronger concern arises when the answer transfers such advice to an unspecified emergency without clearly identifying eligibility, formulation, total dose, and stop conditions. Doubao's inconsistent aspirin milligrams and tablet count exemplify an internal instruction error. Its spray-handling advice can be checked against product-specific labeling, which limits total metered sprays and instructs users not to shake the container. [36] DeepSeek's claim that a sensory effect establishes nitrate freshness similarly exceeds what the product information supports. [35]

Diagnostic certainty can be unsafe even when the suggested tests are correct. An ECG and troponin form part of ACS assessment, but a normal early ECG does not exclude ACS and serial evaluation may be necessary. [1][2][37] Advice that promises a definitive answer within a short interval risks turning access to testing into false reassurance. For special populations, broad cardiac screening schedules also require scrutiny: diabetes guidance does not recommend routine coronary screening in asymptomatic individuals whose risk factors are treated. [38] In bystander resuscitation, clear recognition of unresponsiveness with absent or abnormal breathing and prompt action are more useful than an ambiguous list of triggers. [39]

These mechanisms differ in evidential strength. Some are directly demonstrable contradictions or factual errors; others are conditional risks moderated by appropriate warnings elsewhere in the answer. Preserving this distinction avoids equating every incomplete sentence with proven danger. The binary safety endpoint cannot convey severity, likelihood of implementation, or the protective effect of surrounding advice. The qualitative cases explain why a response deserves concern, while the original majority-vote labels remain the quantitative outcome.

Content quality does not establish accountable patient guidance

Doubao and ChatGPT received similarly favorable DISCERN, EQIP, and GQS ratings, yet neither avoided safety flags or high reading demands. JAMA scores remained low even for ChatGPT, the highest-scoring system on that measure. This is a clinically relevant separation of constructs: organized, useful-looking text can lack verifiable provenance, explicit authorship, disclosure, or currency. JAMA benchmarks identify these transparency features, rather than directly measuring factual correctness. [31] Their absence can make it harder for a clinician or patient to distinguish current evidence from unsupported specificity.

The instruments themselves also impose limits. DISCERN was designed around written treatment choices, EQIP around patient information, and GQS around overall information quality. [28][29][30] Applying them to short conversational answers across prevention, emergency recognition, and recovery requires interpretation. The binary EQIP adaptation further limits direct comparison with studies retaining partial-

credit scoring. High agreement among raters cannot establish that an instrument captures emergency actionability or that a high total offsets a single critical medication error. Source transparency, factual content, and safe actions should consequently remain distinct evaluation targets.

Readability and clinically useful brevity

All models produced text with substantial estimated reading demands, consistent with concerns in earlier MI evaluations. [7][8] Doubao's particularly high grade-level estimates and broad dispersion warrant attention, but extreme formula values should not be read as literal educational prerequisites. Lists, headings, medication names, sentence segmentation, and syllable algorithms can influence the estimates. The lack of the original readability implementation further limits absolute interpretation and reproducibility.

ChatGPT's shorter median answers and favorable content ratings show that greater length was not necessary for higher ratings in this sample. This is descriptive, not evidence that shortening an answer will improve it. For time-critical advice, excessive explanation may compete with the immediate action, while overly compressed language can omit an essential condition. AHRQ's emphasis on easy-to-understand materials and user feedback supports testing whether people can identify the first action, recognize uncertainty, and correctly explain medication instructions. [32] Readability scores should be supplemented by comprehension and action testing with patients and caregivers, including people with limited health literacy.

Implications for clinical practice and research

These outputs are best considered candidate educational text requiring clinical review, particularly around emergency escalation and medication decisions. The evidence does not support autonomous triage or replacement of individualized professional care. A practical review should check that the opening action remains consistent throughout the answer, that doses match the stated formulation and tablet strength, and that a familiar-angina pathway is separated from new suspected infarction. For nurses and rehabilitation teams, the broader domains may help identify recurring questions for reviewed education, but reuse should retain a clear route to professional help.

For model developers and evaluators, the shared flagged questions identify useful challenge areas without proving a common underlying model defect. Future benchmarks should distinguish factual correctness, omission of a necessary condition, contradictory sequencing, and unsupported self-management. Repeated generations and paraphrased prompts should assess how often these problems recur. Direct user studies should measure comprehension, emergency-call timing, medication interpretation, and inappropriate reassurance before claims of improved care are made. A future intervention that enforces a consistent emergency-first structure or checks medication instructions would require prospective testing; this study does not establish its effectiveness.

Strengths and limitations

Strengths include matched comparisons across a broad, balanced question set; three complete independent rating sets; transparent majority-vote and mean aggregation; assessment of both single-rater and average-rater reliability; reporting of all adjusted pairwise comparisons, including nonsignificant results; and safety examples linked to exact outputs and protective context. Question-level bootstrap resampling retained the pairing across models rather than treating outputs or ratings as independent question samples.

Several limitations constrain inference. First, the question set was purposive, English-language, and equally weighted across domains. It cannot represent the distribution of real consultations, regional language needs, complex comorbidity, or evolving dialogue. Excluding individualized or highly contextual questions may also remove difficult real-world use cases. Second, only one output was collected per question and model. Within-model variability, prompt sensitivity, and temporal stability remain unknown; agreement between raters is not test-retest reliability of the chatbots.

Third, models were queried on different days through dynamic web interfaces. Unverified backend snapshots, retrieval settings, account conditions, locale, and session resets limit reproducibility and attribution to a particular model architecture. The use of UK emergency numbers in several English answers despite Chinese investigator affiliations further illustrates the need to document user location and assess localization; this study did not score geographic appropriateness separately. Fourth, ratings were subjective, blinding may have been imperfect, and averaging ordinal single-item scales produces fractional summaries without establishing interval measurement. High ICCs partly reflect variation across questions and models and do not establish instrument validity. Finally, multiple secondary outcomes were evaluated, with adjustment confined to comparisons within each outcome. A nonsignificant safety result in this small matched sample does not demonstrate equivalence or noninferiority; no equivalence margin or safety-powered design was specified. The findings are an exploratory assessment of sampled text, not certification of a chatbot for clinical use.

The reference mapping was not documented as a prospectively locked, independently developed answer key. DISCERN skipped-item handling, EQIP applicability decisions and the original readability implementation could not be independently audited. These gaps limit exact replication and interpretation of the absolute scores. The contextual safety review retained protective statements but did not use a documented independent thematic-coding process. No separate demographic-bias audit or patient comprehension assessment was performed.

Conclusions

Four chatbots answering 70 AMI public questions showed differences in content ratings and readability, while overall safety differences remained statistically inconclusive. Potentially unsafe advice occurred with every model and often coexisted with high accuracy or global quality scores. Emergency instruction

consistency, medication specificity, transparent sourcing, and accessible language warrant evaluation as separate requirements. Repeated-generation testing and patient-based comprehension and behavior studies are needed before inferring safe or effective clinical use.

Declarations

Ethics and consent

This study evaluated publicly generated chatbot responses to standardized questions and did not involve human participants, patient records, or identifiable personal health information. Formal ethics approval was therefore not required.

Privacy and patient involvement

No personal health information or identifiable patient data were entered into any chatbot. All prompts consisted of standardized public health consultation questions.

Competing interests

The authors declare no competing interests.

Use of AI in manuscript preparation

ChatGPT, Gemini, DeepSeek, and Doubao were the systems evaluated in this study, and their generated answers constituted research data. The chatbot systems named in this study were the subjects of evaluation and were used to produce the research responses. Their role is distinct from any AI-assisted writing, editing, translation, reference handling, or figure preparation undertaken during manuscript development.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Author Contribution

M.D. and F.Y. conceived and designed the study. M.D. constructed the 120-item candidate question bank across the seven predefined domains, performed the eligibility screening that reduced it to the final 70 questions, executed all 280 model queries through the four web interfaces on 8–10 June 2026, de-identified and randomly ordered the archived responses before rating, curated the resulting dataset, and

drafted the manuscript. F.Y. supervised the project, secured institutional support, oversaw the rating protocol and the calibration procedure, and critically revised the manuscript. F.S. served as one of the three independent clinical raters, scoring all 280 responses on the seven components, and contributed to the refinement of the safety coding rules and to the resolution of discrepancies during the calibration exercise. Z.W. advised on the statistical analysis plan, verified the Cochran's Q, exact McNemar, Friedman, Wilcoxon and bootstrap procedures, and reviewed the statistical reporting. S.L. contributed to the clinical content review of the question set and to the verification of the extracted response records against the archived transcripts. M.D. and Z.W. produced the tables and figures. All authors discussed the interpretation of the findings, read and approved the final manuscript, and agree to be accountable for all aspects of the work.

Acknowledgments

The authors thank the clinical evaluation team for their contributions to this study.

Data Availability

Supplementary Material 1 provides the CHART checklist; Supplementary Material 2 provides the 120-item screening log and exclusion rationales; Supplementary Material 3 contains the 70 final question texts and all 280 archived question–model responses; Supplementary Material 4 provides the clinical reference mapping; and Supplementary Material 5 provides contextual safety assessments with traceable excerpts. Model-specific summaries, overall tests, primary safety effect estimates and all adjusted pairwise comparisons are reported in the main article. Availability of the individual scoring dataset and analysis code requires author confirmation; no public repository accession is reported.

References

1. Byrne RA, Rossello X, Coughlan JJ, et al. 2023 ESC Guidelines for the management of acute coronary syndromes. Eur Heart J. 2023;44(38):3720–826. 10.1093/eurheartj/ehad191.
2. Rao SV, O'Donoghue ML, Ruel M, et al. 2025 ACC/AHA/ACEP/NAEMSP/SCAI Guideline for the Management of Patients With Acute Coronary Syndromes: A Report of the American College of Cardiology/American Heart Association Joint Committee on Clinical Practice Guidelines. Circulation. 2025;151(13):e771–862. 10.1161/CIR.0000000000001309.
3. Hewett Brumberg EK, Douma MJ, Alibertis K, et al. 2024 American Heart Association and American Red Cross Guidelines for First Aid. Circulation. 2024;150(24):e519–79. 10.1161/CIR.0000000000001281.
4. Chen KC, Hsu CN, Wu CH, et al. 2023 TAMIS/TSOC/TACVPR Consensus Statement for Patients with Acute Myocardial Infarction Rehabilitation. Acta Cardiol Sinica. 2023;39(6):783–806. 10.6515/ACS.202311_39(6).20230921A.

5. Levine GN, Carney RM, Cohen BE, et al. Post-Myocardial Infarction Psychological Distress: A Scientific Statement From the American Heart Association. Circulation. 2025;152(16):e298–310. 10.1161/CIR.0000000000001381.
6. Ayers JW, Poliak A, Dredze M, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. JAMA Intern Med. 2023;183(6):589–96. 10.1001/jamainternmed.2023.1838.
7. Korkmaz I. Accuracy and Reliability of AI Models in Emergency Myocardial Infarction Education. Emerg Med Int. 2026;2026:5530861. 10.1155/emmi/5530861.
8. Lv T, Bao W, Li S, et al. Evaluating large language models for myocardial infarction public health education: a comparative study on information quality, transparency and readability. Front Public Health. 2026;14:1914167. 10.3389/fpubh.2026.1914167.
9. Huo B, Boyle A, Marfo N, et al. Large Language Models for Chatbot Health Advice Studies: A Systematic Review. JAMA Netw Open. 2025;8(2):e2457879. 10.1001/jamanetworkopen.2024.57879.
10. CHART Collaborative. Reporting guideline for chatbot health advice studies: the Chatbot Assessment Reporting Tool (CHART) statement. BMJ Med. 2025;4:e001632. 10.1136/bmjmed-2025-001632.
11. CHART Collaborative. Reporting guidelines for chatbot health advice studies: explanation and elaboration for the Chatbot Assessment Reporting Tool (CHART). BMJ. 2025;390:e083305. 10.1136/bmj-2024-083305.
12. Mese I, Akinci D'Antonoli T, Bluethgen C, et al. Reporting checklist for foundation and large language models in medical research (REFINE): an international consensus guideline. Diagnostic and Interventional Radiology. Published online Febr. 2026;26. 10.4274/dir.2026.263812.
13. Gallifant J, Afshar M, Ameen S, et al. The TRIPOD-LLM reporting guideline for studies using large language models. Nat Med. 2025;31(1):60–9. 10.1038/s41591-024-03425-5.
14. Collet JP, Thiele H, Barbato E, et al. 2020 ESC Guidelines for the management of acute coronary syndromes in patients presenting without persistent ST-segment elevation. Eur Heart J. 2021;42(14):1289–367. 10.1093/eurheartj/ehaa575.
15. Lawton JS, Tamis-Holland JE, Bangalore S, et al. 2021 ACC/AHA/SCAI Guideline for Coronary Artery Revascularization: A Report of the American College of Cardiology/American Heart Association Joint Committee on Clinical Practice Guidelines. Circulation. 2022;145(3):e18–114. 10.1161/CIR.0000000000001038.
16. Jacobs AK, Ali MJ, Best PJ, et al. Systems of Care for ST-Segment-Elevation Myocardial Infarction: A Policy Statement From the American Heart Association. Circulation. 2021;144(20):e310–27. 10.1161/CIR.0000000000001025.
17. Tamis-Holland JE, Abbott JD, Al-Azizi K, et al. SCAI Expert Consensus Statement on the Management of Patients With STEMI Referred for Primary PCI. J Soc Cardiovasc Angiography Interventions. 2024;3(11):102294. 10.1016/j.jscai.2024.102294.

18. Kim HK, Ryoo S, Lee SH, et al. 2024 Korean Society of Myocardial Infarction/National Evidence-Based Healthcare Collaborating Agency Guideline for the Pharmacotherapy of Acute Coronary Syndromes. Korean Circulation J. 2024;54(12):767–93. 10.4070/kcj.2024.0257.
19. Welt FGP, Batchelor W, Spears JR, et al. Reperfusion Injury in Patients With Acute Myocardial Infarction: JACC Scientific Statement. J Am Coll Cardiol. 2024;83(22):2196–213. 10.1016/j.jacc.2024.02.056.
20. Damluji AA, van Diepen S, Katz JN, et al. Mechanical Complications of Acute Myocardial Infarction: A Scientific Statement From the American Heart Association. Circulation. 2021;144(2):e16–35. 10.1161/CIR.0000000000000985.
21. Ozaki Y, Hara H, Onuma Y, et al. CVIT expert consensus document on primary percutaneous coronary intervention (PCI) for acute myocardial infarction (AMI) update 2022. Cardiovasc Intervention Ther. 2022;37(1):1–34. 10.1007/s12928-021-00829-9.
22. Schlotter F, Huber K, Hassager C, et al. Ventricular septal defect complicating acute myocardial infarction: diagnosis and management. A Clinical Consensus Statement of the Association for Acute CardioVascular Care (ACVC) of the ESC, the European Association of Percutaneous Cardiovascular Interventions (EAPCI) of the ESC and the ESC Working Group on Cardiovascular Surgery. Eur Heart J. 2024;45(28):2478–92. 10.1093/eurheartj/ehae363.
23. Kaczorowski DJ, Takeda K, Atluri P, et al. 2025 American Association for Thoracic Surgery (AATS) Expert Consensus Document: Surgical management of acute myocardial infarction and associated complications. J Thorac Cardiovasc Surg. 2025;170(5):1327–44. 10.1016/j.jtcvs.2025.04.013.
24. OpenAI. Introducing GPT-5.5. April 23. 2026. Accessed November 14, 2026. https://openai.com/index/introducing-gpt-5-5/
25. Google. Gemini 3.1 Pro: A smarter model for your most complex tasks. February 19, 2026. Accessed November 14, 2026. https://blog.google/innovation-and-ai/models-and-research/gemini-models/gemini-3-1-pro/
26. DeepSeek. Introducing DeepSeek-V3.2-Exp. September 29, 2025. Accessed November 14, 2026. https://www.deepseek.com/en/news/v3-2-exp/
27. ByteDance Seed. Seed 2.0 Official Launch. February 14. 2026. Accessed November 14, 2026. https://seed.bytedance.com/en/blog/seed-2-0-official-launch
28. Charnock D, Shepperd S, Needham G, et al. DISCERN: an instrument for judging the quality of written consumer health information on treatment choices. J Epidemiol Commun Health. 1999;53(2):105–11. 10.1136/jech.53.2.105.
29. Moult B, Franck LS, Brady H. Ensuring Quality Information for Patients: development and preliminary validation of a new instrument to improve the quality of written health care information. Health Expect. 2004;7(2):165–75. 10.1111/j.1369-7625.2004.00273.x.
30. Bernard A, Langille M, Hughes S, et al. A systematic review of patient inflammatory bowel disease information resources on the World Wide Web. Am J Gastroenterol. 2007;102(9):2070–7. 10.1111/j.1572-0241.2007.01325.x.

31. Silberg WM, Lundberg GD, Musacchio RA. Assessing, controlling, and assuring the quality of medical information on the Internet: caveant lector et viewor. JAMA. 1997;277(15):1244–5. 10.1001/jama.1997.03540390074039.
32. Agency for Healthcare Research and Quality. Health Literacy Universal Precautions Toolkit, 3rd Edition. Tool 11: Assess, Select, and Create Easy-to-Understand Materials. 2024. Accessed November 14, 2026. https://www.ahrq.gov/health-literacy/improve/precautions/tool11.html
33. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J Royal Stat Soc Ser B (Methodological). 1995;57(1):289–300. 10.1111/j.2517-6161.1995.tb02031.x.
34. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J Chiropr Med. 2016;15(2):155–63. 10.1016/j.jcm.2016.02.012.
35. Viatris Specialty LLC. NITROSTAT (nitroglycerin) sublingual tablets. Prescribing information and patient information. DailyMed. Updated February 15, 2023. Accessed November 14, 2026. https://dailymed.nlm.nih.gov/dailymed/drugInfo.cfm?setid=9a52da75-d0af-4237-96c0-0b12adfa8636
36. Independence Pharmaceuticals LLC. NITROLINGUAL PUMPSPRAY (nitroglycerin) lingual spray. Prescribing information and instructions for use. DailyMed. Updated May 13, 2024. Accessed November 14, 2026. https://dailymed.nlm.nih.gov/dailymed/drugInfo.cfm?setid=d19b5f51-ca85-48d6-9ccd-8ae26249cdc3
37. National Institute for Health and Care Excellence. Recent-onset chest pain of suspected cardiac origin: assessment and diagnosis. Clinical guideline CG95. Published March 24, 2010; updated November 30, 2016. Accessed November 14. 2026. https://www.nice.org.uk/guidance/cg95/chapter/recommendations
38. American Diabetes Association Professional Practice Committee for Diabetes. 10. Cardiovascular disease and risk management: Standards of Care in Diabetes-2026. Diabetes Care. 2026;49(suppl 1):S216–S245. 10.2337/dc26-S010
39. Resuscitation Council UK. Adult basic life support Guidelines. Resuscitation Guidelines. 2025. Accessed November 14, 2026. https://www.resus.org.uk/professional-library/2025-resuscitation-guidelines/adult-basic-life-support-guidelines

Figures

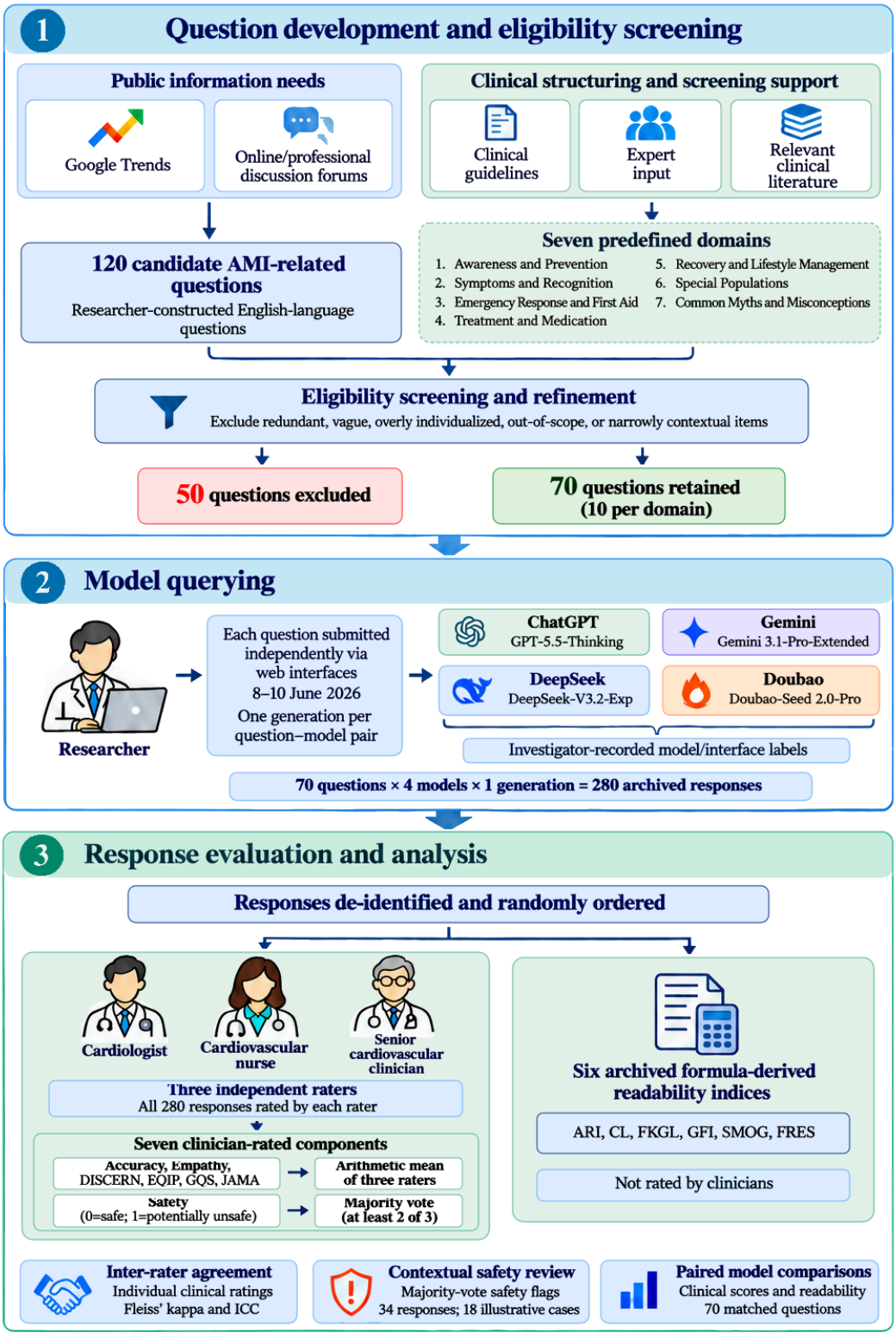


Figure 1

Study workflow



Figure 2

Safety accuracy and empathy across chatbots

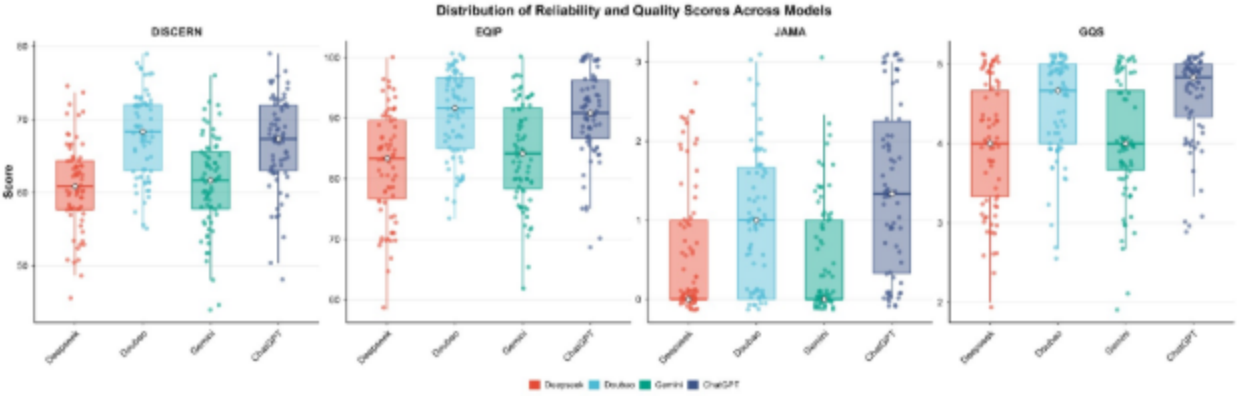


Figure 3

Information quality and transparency

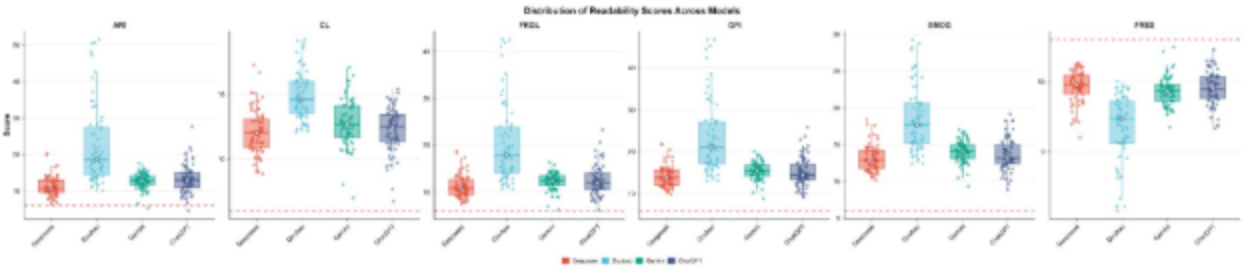


Figure 4

Archived readability estimates

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryMaterial1CHARTChecklist.docx
- SupplementaryMaterial5QualitativeAnalysisofUnsafeResponses.docx
- SupplementaryMaterial4GuideMapping.docx
- SupplementaryMaterial2ScreeningProcess.docx
- SupplementaryMaterial3AnswerRecords.docx