

Evaluating the Readability, Quality, and Reliability of Responses Generated by ChatGPT, Gemini, and Perplexity to the Most Frequently Searched Queries About Acute Mesenteric Ischemia

✉ Sema Ayten¹, ✉ Nihal Dolanbay²

¹Göztepe Prof. Dr. Süleyman Yalçın City Hospital, Clinic of Emergency Medicine, İstanbul, Türkiye

²SANKO University Faculty of Health Sciences, Department of Surgical Nursing, Gaziantep, Türkiye

Abstract

Aim: Patients increasingly consult artificial intelligence (AI) chatbots before deciding whether to seek emergency care. This study compared the readability, reliability, and quality of responses generated by ChatGPT, Gemini, and Perplexity in response to frequently searched queries about acute mesenteric ischemia.

Materials and Methods: Sixteen queries identified through Google Trends were submitted to each chatbot, yielding 48 responses. Readability was assessed with seven formulas. Two evaluators independently rated reliability using the JAMA benchmarks and modified DISCERN and quality using the Global Quality score and the Ensuring Quality Information for Patients tool. Platforms were compared with the Friedman and Wilcoxon signed-rank tests.

Results: ChatGPT produced shorter responses than Gemini and Perplexity (medians 180.5, 390.0, and 416.5 words; $p<0.001$). All platforms' content was written at a college reading level (median Flesch-Kincaid Grade Level of 14.0-15.5), well above the recommended sixth to eighth-grade level. Perplexity obtained the highest JAMA and Ensuring Quality Information for Patients scores, whereas Gemini and Perplexity outperformed ChatGPT on the modified DISCERN and the Global Quality score.

Conclusion: No chatbot was superior across all measures, and none produced text readable by the average patient. Information on this time-critical emergency should be simplified and professionally reviewed before being recommended to patients. The clinical accuracy and the appropriateness of emergency referral advice were not assessed and require further evaluation.

Keywords: Mesenteric ischemia, artificial intelligence, chatbots, health literacy, readability

Introduction

Acute mesenteric ischemia (AMI; not to be confused with acute myocardial infarction) is an uncommon but highly lethal abdominal emergency in which an abrupt reduction in intestinal blood flow leads, if untreated, to bowel necrosis, multiorgan failure, and death (1-3). The condition comprises arterial embolism, arterial thrombosis, mesenteric venous thrombosis, and non-occlusive mesenteric ischemia, because each follows a different clinical course, the presentation is often nonspecific and

the diagnosis is frequently delayed (3-5). No laboratory biomarker is accurate enough to establish the diagnosis on its own (6), and computed tomography angiography remains the cornerstone of evaluation (3,7). Management ranges from endovascular or open revascularization to bowel resection, depending on the mechanism and bowel viability (1,8,9), and every delay increases bowel loss and mortality (3).

Artificial intelligence (AI) has so far been explored in AMI mainly as a clinician-facing tool, for example to predict mortality or to interpret computed tomography angiography images (4,10,11).



Corresponding Author: Sema Ayten MD, Göztepe Prof. Dr. Süleyman Yalçın City Hospital, Clinic of Emergency Medicine, İstanbul, Türkiye

E-mail: semayten@hotmail.com **ORCID ID:** orcid.org/0000-0001-8417-331X

Cite this article as: Ayten S, Dolanbay N. Evaluating the readability, quality, and reliability of responses generated by ChatGPT, Gemini, and Perplexity to the most frequently searched queries about acute mesenteric ischemia. Eurasian J Emerg Med. 2026;25: 494-501

Received: 28.09.2026

Accepted: 15.10.2026

Published: 27.10.2026



©Copyright 2026 The Author(s). The Emergency Physicians Association of Turkey / Eurasian Journal of Emergency Medicine published by Galenos Publishing House. Licensed by Creative Commons Attribution-NonCommercial-NoDerivatives (CC BY-NC-ND) 4.0 International License

A different and far less-studied application is the patient-facing chatbot. Large language model-based platforms such as ChatGPT, Gemini, and Perplexity now answer health questions in conversational language and are increasingly evaluated as sources of medical information (12-14). Their answers, however, vary considerably in reliability and readability, and some appear convincing while conflicting with clinical practice guidelines (12-15).

Because patients may consult these platforms before seeking professional care (16), the information they provide must be understandable, reliable, and clinically appropriate. Studies across several specialties have shown that chatbot responses frequently exceed the reading level recommended for patient materials (14,17-19). Structured instruments allow these characteristics to be assessed systematically: the JAMA benchmarks examine authorship, attribution, disclosure, and currency (20); DISCERN evaluates the reliability of consumer health information (21); the Global Quality score (GQS) rates overall quality and usefulness (22); and the Ensuring Quality Information for Patients (EQIP) tool assesses the content, identification, and structure of written patient information (23).

For a time-critical emergency such as AMI, the practical question is not only whether chatbot information is readable and reliable, but also whether it helps patients recognize that urgent evaluation is needed. To our knowledge, chatbot-generated information on AMI has not been evaluated previously.

This study aimed to compare the readability, reliability, and quality of responses generated by ChatGPT, Gemini, and Perplexity in response to the most frequently searched queries about AMI.

Materials and Methods

Study Design

This comparative cross-sectional study evaluated the information generated by three AI platforms. The evaluated platforms were ChatGPT, Gemini, and Perplexity. The study focused on patient-oriented information about AMI. No patients, medical records, or personally identifiable information were included. The study was reported in accordance with the Chatbot Assessment Reporting Tool guideline, where applicable (24).

Selection of Search Queries

Frequent search queries related to AMI were identified using Google Trends (Google LLC, Mountain View, CA, USA) on 10 July

2026. The search was conducted using “AMI” as the initial search term, covering the preceding five years, with the geographical setting worldwide, without country-specific restrictions. Related queries were examined under the “Top” category, and the 16 highest-ranked eligible queries were selected for analysis. Duplicate or overlapping queries and those without a plausible clinical relationship to AMI were excluded. Queries were retained in their original keyword format rather than converted into questions to reflect actual public search behavior, consistent with previous studies (12,14).

Broader terms, including “Thrombosis,” “Acute Abdomen,” and “Peritonitis,” were retained because they appeared among the top-ranked related queries and represented clinically relevant concepts associated with the vascular mechanisms, clinical presentation, or complications of AMI. Although these terms are not specific to AMI, their inclusion was intended to reflect broader information-seeking behavior related to the condition. All three platforms interpreted the abbreviated query “SMA” in the context of the superior mesenteric artery.

1. Acute mesenteric ischemia
2. Mesenteric ischemia treatment
3. Mesenteric ischemia symptoms
4. Mesenteric artery
5. Acute mesenteric ischemia symptoms
6. Thrombosis
7. Superior mesenteric artery
8. SMA
9. Acute abdomen
10. Intestinal ischemia
11. Peritonitis
12. Mesenteric ischemia radiology
13. Mesenteric venous thrombosis
14. Bowel infarction
15. Bowel necrosis
16. Mesenteric ischemia diagnosis

Artificial Intelligence Response Collection

Responses were collected on 10 July 2026 through the official web interfaces of ChatGPT (OpenAI, San Francisco, CA, USA), Gemini (Google LLC, Mountain View, CA, USA), and Perplexity (Perplexity AI, San Francisco, CA, USA). All three platforms were accessed using their freely available versions without requiring user login.

These platforms use proprietary, closed-source models. Default settings were maintained, and no specific model was manually selected. Because the interfaces did not provide independently verifiable model identifiers for each response, the exact underlying model versions could not be confirmed. Perplexity was accessed using its default automatic model-selection mode.

Each of the 16 queries was entered in English, exactly as retrieved from Google Trends, in a separate, new chat session without prior context, additional instructions, or prompt modification. Only the first response generated for each query was recorded, yielding 48 responses for analysis. No repeated queries or follow-up prompts were used.

The sample size was determined by the 16 highest-ranked eligible queries identified through Google Trends, rather than by a formal power calculation. Each query was submitted once to each of the three platforms, resulting in 48 responses. This approach was intended to provide a standardized cross-sectional comparison of the platforms under identical query conditions at a single point in time, rather than to assess within-platform response variability.

Readability Assessment

Readability was calculated using the Readability Formulas online tool (<https://readabilityformulas.com>). Seven measures were computed for each response: the Flesch Reading Ease score (FRES) (25), Flesch-Kincaid Grade Level (FKGL) (26), Gunning Fog index (27), Simple Measure of Gobbledygook (SMOG) (28), Automated Readability index (ARI) (29), Coleman-Liau index (30), and Linsear Write Formula. FRES ranges from 0 to 100, with higher scores indicating easier text; the remaining measures estimate the United States school grade level required to understand the text. A sixth- to eighth-grade level is generally recommended for patient education materials (31). Before readability analysis, all chatbot responses underwent standardized text preprocessing. URLs, citations, reference markers, headings, bullet-point symbols, and other non-prose elements were removed. The remaining textual content was analyzed using the seven readability formulas. The same preprocessing procedure was applied consistently across all three platforms to minimize potential formatting-related differences in readability scores.

Quality and Reliability Assessment

Reliability was assessed with the JAMA benchmarks (20) and the modified DISCERN (21,32). The JAMA benchmarks award one point each for authorship, attribution, disclosure, and currency (total score 0-4). The modified DISCERN consists of five yes/no questions addressing clarity, reliability of sources,

balance, provision of additional sources, and reference to areas of uncertainty (total score 0-5). Overall quality was rated using the GQS, a five-point scale ranging from 1 (poor) to 5 (excellent) (22). Patient information quality was assessed with the EQIP tool (23). The original 20-item version (EQIP-20) was used: each item was rated as “yes”, “partly”, “no”, or “not applicable”, and the total score was expressed as a percentage, with higher scores indicating better quality.

All responses were independently rated using the four instruments by two evaluators with different professional backgrounds: an emergency medicine specialist and a faculty member holding a PhD in surgical nursing. Because Perplexity displays source links within its responses, complete blinding to the platform was not feasible. Inter-rater agreement was calculated from the independent ratings; disagreements were then resolved by discussion, and the consensus scores were used in all comparative analyses.

Statistical Analysis

Statistical analyses were performed using IBM SPSS Statistics for Windows, version 27.0 (IBM Corp., Armonk, NY, USA). Continuous variables are presented as median (minimum–maximum) and 25th–75th percentiles. Differences among the three platforms were assessed using the Friedman test, and pairwise comparisons were evaluated using the Wilcoxon signed-rank test. A Bonferroni correction was applied to the three pairwise comparisons; therefore, $p < 0.0167$ was considered significant for these tests, while $p < 0.05$ was considered significant for other tests. Effect sizes for the Friedman test were expressed as Kendall's W , calculated as $\chi^2/[N(k-1)]$, where N is the number of queries and k the number of platforms; values of approximately 0.1, 0.3, and 0.5 were interpreted as small, moderate, and large effects, respectively.

Inter-rater reliability was evaluated using intraclass correlation coefficients (ICCs) based on a two-way mixed-effects model with absolute agreement; single-measure ICC values with 95% confidence intervals (CIs) are reported. ICC values below 0.50, between 0.50 and 0.75, between 0.75 and 0.90, and above 0.90 were interpreted as poor, moderate, good, and excellent agreement, respectively (33).

Results

A total of 48 responses from three AI platforms were evaluated. Each platform generated responses to the same 16 AMI-related queries. Effect sizes for the overall comparisons ranged from moderate (Kendall's $W = 0.31$ for Linsear Write) to large (Kendall's $W = 1.00$ for JAMA), and are presented in Table 1.

Table 1. Overall comparison of readability, reliability, and quality measures among ChatGPT, Gemini, and Perplexity

Parameter	ChatGPT mean rank	Gemini mean rank	Perplexity mean rank	χ^2	df	p	Kendall's W
Word count	1.13	2.38	2.50	18.500	2	<0.001	0.58
FRES	1.06	2.78	2.16	25.410	2	<0.001	0.79
FKGL	2.16	1.31	2.53	12.667	2	0.002	0.40
Gunning Fog index	2.75	1.19	2.06	19.625	2	<0.001	0.61
SMOG	1.13	2.00	2.88	24.500	2	<0.001	0.77
ARI	1.44	1.56	3.00	24.125	2	<0.001	0.75
Coleman-Liau index	2.00	1.38	2.63	12.500	2	0.002	0.39
Linsear Write Formula	2.63	1.81	1.56	9.875	2	0.007	0.31
JAMA	1.50	1.50	3.00	32.000	2	<0.001	1.00
Modified DISCERN	1.22	2.47	2.31	19.000	2	<0.001	0.59
GQS	1.50	2.16	2.34	12.182	2	0.002	0.38
EQIP	1.28	2.16	2.56	15.404	2	<0.001	0.48

Friedman test. χ^2 : chi-square, df: degrees of freedom, FRES: Flesch Reading Ease Score, FKGL: Flesch-Kincaid Grade Level, SMOG: Simple Measure of Gobbledygook, ARI: Automated Readability index, JAMA: Journal of the American Medical Association benchmarks, GQS: Global Quality score, EQIP: Ensuring Quality Information for Patients. Kendall's $W = \chi^2 / [N(k-1)]$, N=16 queries, k=3 platforms

ChatGPT produced a median word count of 180.50 words. Median word counts were 390.00 for Gemini and 416.50 for Perplexity. Overall word counts differed significantly among the platforms ($p < 0.001$). ChatGPT responses were shorter than Gemini responses ($p = 0.001$). They were also shorter than Perplexity responses ($p = 0.001$). Gemini and Perplexity showed no significant difference ($p = 0.877$).

Median FRES scores were 0.00 for ChatGPT, 16.50 for Gemini, and 9.00 for Perplexity. FRES differed significantly among the three platforms ($p < 0.001$). ChatGPT had lower FRES scores than Gemini ($p < 0.001$). ChatGPT also had lower scores than Perplexity ($p = 0.001$). Gemini had higher FRES scores than Perplexity ($p = 0.014$).

FKGL scores differed significantly among the platforms ($p = 0.002$). ChatGPT had higher FKGL scores than Gemini ($p = 0.013$). ChatGPT and Perplexity showed no significant difference ($p = 0.074$). Gemini had lower FKGL scores than Perplexity ($p = 0.010$). The median FKGL values for the three platforms (15.52, 14.04, and 15.47 for ChatGPT, Gemini, and Perplexity) corresponded to a college reading level, exceeding the recommended sixth- to eighth-grade level.

Gunning Fog scores differed significantly among the platforms ($p < 0.001$). ChatGPT scored higher than Gemini ($p = 0.001$). ChatGPT also had higher scores than Perplexity ($p = 0.006$). Gemini had lower scores than Perplexity ($p = 0.010$).

SMOG scores differed significantly among the three platforms ($p < 0.001$). ChatGPT had lower SMOG scores than Gemini ($p = 0.001$). ChatGPT also scored lower than Perplexity ($p < 0.001$). Gemini had lower scores than Perplexity ($p < 0.001$).

ARI scores also differed significantly among the platforms ($p < 0.001$). Perplexity's scores were higher than those of ChatGPT and Gemini (both $p < 0.001$). ChatGPT and Gemini showed no significant difference ($p = 0.698$).

Coleman-Liau scores differed significantly among the platforms ($p = 0.002$). Gemini had significantly lower scores than Perplexity ($p = 0.004$). Other pairwise comparisons remained nonsignificant after Bonferroni correction. Linsear Write scores also differed significantly among platforms ($p = 0.007$). ChatGPT had higher scores than Perplexity ($p = 0.003$). Other pairwise comparisons were not significant after correction. Detailed readability findings are presented in Tables 1 and 2.

JAMA scores differed significantly among the three platforms ($p < 0.001$). ChatGPT and Gemini both had median JAMA scores of 0.00. Perplexity had a median JAMA score of 2.00. Perplexity outperformed both competing platforms ($p < 0.001$ for both comparisons). ChatGPT and Gemini showed no difference ($p = 1.000$).

Modified DISCERN scores also differed significantly among platforms ($p < 0.001$). Median scores were 2.50 for ChatGPT, 4.00 for Gemini, and 3.00 for Perplexity. Gemini scored higher than ChatGPT ($p = 0.001$). Perplexity also scored higher than

ChatGPT ($p=0.001$). Gemini and Perplexity showed no significant difference ($p=0.527$).

GQS scores differed significantly among the three platforms($p=0.002$). Median scores were 3.00 for ChatGPT and 4.00 for both Gemini and Perplexity. Gemini scored higher than ChatGPT ($p=0.008$). Perplexity also scored higher than ChatGPT ($p=0.003$). Gemini and Perplexity showed no significant difference ($p=0.414$).

EQIP scores differed significantly among the three platforms ($p<0.001$). Median scores were 47.00 for ChatGPT, 49.00 for Gemini, and 52.00 for Perplexity. Gemini scored higher than GPT ($p=0.002$). Perplexity also scored higher than ChatGPT ($p=0.001$). Perplexity scored higher than Gemini ($p=0.013$). These findings are presented in Tables 1 and 3.

Inter-rater agreement varied across the four instruments. The two raters assigned identical JAMA scores to all 48 responses (100% agreement; ICC=1.000, 95% CI: 1.000-1.000), and EQIP showed good agreement (ICC=0.900, 95% CI: 0.857-0.931). For modified DISCERN (ICC=0.679, 95% CI: 0.272–0.880) and GQS (ICC=0.638, 95% CI: 0.170-0.874), the point estimates indicated moderate agreement, although the wide CIs ranged from poor to good (Table 4).

Discussion

In this study, the three chatbots differed in almost every measure, but none performed best across all of them. Perplexity scored highest on the JAMA benchmarks and EQIP; Gemini and Perplexity scored higher than ChatGPT on modified DISCERN

Parameter	ChatGPT median (min-max) (25 th -75 th)	Gemini median (min-max) (25 th -75 th)	Perplexity median (min-max) (25 th -75 th)	ChatGPT vs. Gemini, p	ChatGPT vs. Perplexity, p	Gemini vs. Perplexity, p
Word count	180.50 (113-496) [138.50-206.75]	390.00 (211-617) [322.25-458.25]	416.50 (299-576) [324.25-469.00]	0.001	0.001	0.877
FRES	0.00 (0-20) [0.00-5.25]	16.50 (4-56) [9.00-32.00]	9.00 (0-35) [2.75-18.25]	<0.001	0.001	0.014
FKGL	15.52 (12-16) [13.91-15.82]	14.04 (8-21) [11.56-15.16]	15.47 (13-18) [14.87-16.19]	0.013	0.074	0.010
Gunning Fog index	18.25 (16-20) [16.93-19.18]	16.65 (10-18) [14.28-17.38]	17.45 (15-19) [16.50-18.20]	0.001	0.006	0.010
SMOG	8.08 (7-11) [7.74-8.61]	10.81 (8-13) [10.34-11.78]	12.92 (12-14) [12.19-13.44]	0.001	<0.001	<0.001
ARI	13.35 (9-15) [12.60-14.17]	13.35 (9-16) [11.26-14.67]	16.43 (12-19) [14.72-16.82]	0.698	<0.001	<0.001
Coleman-Liau index	17.35 (12-19) [16.36-18.36]	16.07 (11-19) [14.02-17.55]	17.81 (12-21) [15.64-19.07]	0.063	0.171	0.004
Linsear Write Formula	18.34 (11-39) [15.60-28.94]	14.16 (9-23) [12.81-17.12]	13.32 (10-23) [12.14-14.79]	0.017	0.003	0.326

Data are presented as median (minimum-maximum) and [25th-75th percentiles]. Pairwise comparisons: Wilcoxon signed-rank test with Bonferroni correction ($p<0.0167$ considered significant). FRES: Flesch Reading Ease Score, FKGL: Flesch-Kincaid Grade Level, SMOG: Simple Measure of Gobbledygook, ARI: Automated Readability index, min-max: minimum-maximum

Parameter	ChatGPT median (min-max) [25 th -75 th]	Gemini median (min-max) [25 th -75 th]	Perplexity median (min-max) [25 th -75 th]	ChatGPT vs. Gemini, p	ChatGPT vs. Perplexity, p	Gemini vs. Perplexity, p
JAMA	0.00 (0-0) [0.00-0.00]	0.00 (0-0) [0.00-0.00]	2.00 (2-2) [2.00-2.00]	1.000	<0.001	<0.001
Modified DISCERN	2.50 (2-3) [2.00-3.00]	4.00 (3-4) [3.00-4.00]	3.00 (3-4) [3.00-4.00]	0.001	0.001	0.527
GQS	3.00 (3-4) [3.00-4.00]	4.00 (3-4) [3.25-4.00]	4.00 (3-4) [4.00-4.00]	0.008	0.003	0.414
EQIP	47.00 (45-48) [46.00-47.75]	49.00 (45-55) [48.00-50.00]	52.00 (45-56) [49.25-54.00]	0.002	0.001	0.013

Data are presented as median (minimum-maximum) [25th-75th percentiles]. Pairwise comparisons: Wilcoxon signed-rank test with Bonferroni correction ($p<0.0167$ considered significant). JAMA: Journal of the American Medical Association benchmarks, DISCERN: Modified DISCERN, GQS: Global Quality score, EQIP: Ensuring Quality Information for Patients: min-max: minimum-maximum

Table 4. Inter-rater reliability of the quality and reliability assessment tools			
Scoring system	ICC	95% CI	Interpretation
JAMA	1.000	1.000-1.000	Excellent
Modified DISCERN	0.679	0.272-0.880	Moderate
GQS	0.638	0.170-0.874	Moderate
EQIP	0.900	0.857-0.931	Good
Two-way mixed-effects model, absolute agreement, single measures. JAMA scores did not vary within platforms; therefore, the ICC of 1.000 reflects identical ratings on explicit criteria. The wide confidence intervals for modified DISCERN and GQS extend from poor to good agreement. ICC: Interclass correlation coefficient, CI: Confidence interval, GQS: Global Quality score, EQIP: Ensuring Quality Information for Patients			

and GQS; ChatGPT produced the shortest responses. All three platforms generated text at a college reading level, which is well above the level recommended for patient information.

Patient education materials are recommended to be written at a sixth- to eighth-grade reading level (31). In our study, median FKGL values ranged from 14.04 to 15.52, and median FRES values between 0 and 16.5 fall within the range classified as very difficult to read (25). This finding is consistent with previous chatbot studies on cardiopulmonary resuscitation, sexually transmitted diseases, and prostate cancer radiotherapy, in which responses also exceeded recommended levels (14,17,18). For AMI, the gap is particularly relevant: a patient with sudden, severe abdominal pain who receives complex, terminology-heavy text may fail to extract the single most important message: that emergency evaluation is needed.

Pairwise readability comparisons depended on the formula used. ChatGPT had the lowest FRES, indicating the most difficult text, yet it also had the lowest SMOG score. This apparent contradiction is more likely to reflect properties of the formulas than a true difference in readability. SMOG was developed for samples of 30 sentences (28), whereas ChatGPT responses were considerably shorter (a median of 180.5 words). Although URLs, citations, reference markers, and other non-prose elements were removed before analysis, differences in vocabulary, sentence structure, and response length may still have contributed to variations among readability formulas. Individual formula results should therefore be interpreted cautiously; the consistent message across formulas is that no platform approached the recommended level.

ChatGPT generated substantially shorter responses than the other two platforms, which may reflect different default response strategies (12-14). Longer answers allow broader explanations, but length alone does not improve understanding and may increase the cognitive burden for a patient with acute symptoms.

Perplexity was the only platform to obtain JAMA points, reflecting its practice of displaying source links alongside responses, which corresponds to the attribution and currency criteria of the JAMA benchmarks (20). Similar advantages of Perplexity have been reported in previous comparisons (13,14). However, the presence of citations does not guarantee that the cited sources are accurate, relevant, or even existent. None of the platforms provided authorship or disclosure information.

Gemini and Perplexity outperformed ChatGPT on modified DISCERN, GQS, and EQIP; however, they did not differ from each other on modified DISCERN and GQS. Similarly, a previous study evaluating cardiopulmonary resuscitation training videos on YouTube demonstrated considerable variability in the quality and reliability of digital health information, highlighting the importance of standardized assessment tools and professional oversight (34). These instruments reward balance, discussion of uncertainty, structure, and completeness (21-23), features more likely to be present in longer responses. This may explain why the two platforms with longer responses scored higher, but it also illustrates a trade-off: more complete responses were not easier to read. Comparable platform-related differences have been reported for palliative care, cardiopulmonary resuscitation, microtia, and breast cancer surgery (13,14,19,35).

These findings have specific implications for emergency medicine. Mortality in AMI increases with delays in diagnosis and revascularization (2,5), and structured management pathways aim to shorten this interval (1,8). A patient-facing chatbot is useful in this setting only if it clearly conveys that sudden, severe abdominal pain, particularly in older patients or in those with atrial fibrillation or vascular disease, requires immediate emergency evaluation (1,3). The instruments used in this study do not specifically assess the presence of such urgency messages, and future studies should evaluate this outcome directly. Importantly, higher reliability or quality scores should not be interpreted as evidence of clinical accuracy or guideline adherence. In the context of AMI, the absence of a direct assessment of emergency referral advice and of guideline-consistent recommendations limits the clinical interpretation of these findings.

The AI tools studied in AMI to date, such as mortality prediction models and computed tomography angiography analysis (4,10,11), are designed to support clinicians. Patient-facing chatbots differ fundamentally because their output reaches users without professional filtering. Evidence that chatbot recommendations may conflict with clinical guidelines (15) and that their quality varies across platforms (35) underlines the need for professional oversight.

Although the ICC point estimates indicated moderate inter-rater agreement for modified DISCERN and GQS, the wide 95% CIs suggest considerable uncertainty regarding the precision of these estimates. Therefore, differences between platforms based on these two instruments should be interpreted cautiously, as variability between evaluators may have influenced the observed results. The different professional backgrounds of the two evaluators, emergency medicine and surgical nursing, may also have contributed to the lower agreement for these more subjective instruments, although this combination reflects both the clinical and patient-education perspectives.

Study Limitations

This study has several limitations. First, only three platforms were evaluated; their responses may change with model updates. Second, only one response per query was evaluated. Given the stochastic nature of generative AI, repeated queries may yield different responses. This approach limits the assessment of within-platform variability, reproducibility, and generalizability. Future studies should incorporate repeated queries across multiple time points to evaluate response consistency and stability more comprehensively. Third, queries were entered in the keyword format retrieved from Google Trends. Although this reflects search-engine behavior and is consistent with previous studies (12,14), it may not represent conversational, question-based interactions, which could produce different answers. Some queries were broader terms related to AMI rather than specific to it. Although these broader queries were retained to reflect real-world search behavior, their inclusion may have influenced the observed differences in chatbot performance. Future studies focusing exclusively on AMI-specific queries could help determine whether the present findings remain consistent when broader or ambiguous search terms are excluded. Fourth, only English-language queries were analyzed. Fifth, readability formulas measure linguistic complexity but not actual patient comprehension. Sixth, the quality instruments include subjective components; the two evaluators were from different professional backgrounds; complete blinding of the evaluators to the platform was not feasible because Perplexity displays its sources. Seventh, because the platforms were accessed through free, non-logged-in web interfaces, the exact underlying model versions could not be independently verified; the findings should therefore be interpreted as comparisons of the publicly accessible chatbot platforms under the specified access conditions rather than as comparisons of definitively identified individual language models. Clinical accuracy and guideline adherence were not directly evaluated, including whether responses appropriately recommended an emergency assessment for suspected AMI. Therefore, higher quality or reliability scores should

not be interpreted as evidence of clinically accurate or safe recommendations. Future studies should include larger, question-based query sets, repeated sampling, and direct, guideline-based assessment of clinical accuracy and emergency referral advice.

Conclusion

ChatGPT, Gemini, and Perplexity differed in the readability, reliability, and quality of their responses to frequently searched AMI queries, and no platform was superior across all measures. Perplexity provided the most transparent sourcing and the highest quality of patient information, whereas Gemini and Perplexity outperformed ChatGPT in reliability and overall quality. All three platforms, however, produced text well above the recommended reading level. Given the time-critical nature of AMI, these findings highlight the need for more accessible patient information. However, the clinical accuracy, guideline adherence, and appropriateness of emergency referral advice provided by these platforms require further evaluation before their use as reliable sources of patient guidance can be recommended.

Ethics

Ethics Committee Approval: This study evaluated publicly accessible responses generated by artificial intelligence platforms. No human participants or identifiable personal data were included. Therefore, institutional ethics approval was not required.

Informed Consent: This comparative cross-sectional study evaluated the information generated by three AI platforms. The evaluated platforms were ChatGPT, Gemini, and Perplexity. The study focused on patient-oriented information about AMI. No patients, medical records, or personally identifiable information were included.

Footnotes

Authorship Contributions

Concept: S.A., N.D., Design: S.A., N.D., Data Collection or Processing: S.A., N.D., Analysis or Interpretation: S.A., N.D., Literature Search: S.A., N.D., Writing: S.A., N.D.

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

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

References

1. Bala M, Catena F, Kashuk J, De Simone B, Gomes CA, Weber D, et al. Acute mesenteric ischemia: updated guidelines of the World Society of Emergency Surgery. *World J Emerg Surg.* 2022;17:54.

2. Tamme K, Reintam Blaser A, Laisaar KT, Mändul M, Kals J, Forbes A, et al. Incidence and outcomes of acute mesenteric ischaemia: a systematic review and meta-analysis. *BMJ Open*. 2022;12:e062846.
3. Garzelli L, Ben Abdallah I, Nuzzo A, Zappa M, Corcos O, Dioguardi Burgio M, et al. Insights into acute mesenteric ischaemia: an up-to-date, evidence-based review from a mesenteric stroke centre unit. *Br J Radiol*. 2023;96:20230232.
4. Bourcier S, Klug J, Nguyen LS. Non-occlusive mesenteric ischemia: diagnostic challenges and perspectives in the era of artificial intelligence. *World J Gastroenterol*. 2021;27:4088-103.
5. Martini V, Lederer AK, Fink J, Chikhladze S, Utzolino S, Fichtner-Feigl S, et al. CLINEX guidelines: acute mesenteric ischaemia. *Langenbecks Arch Surg*. 2022;407:1225-32.
6. Reintam Blaser A, Starkopf J, Björck M, Forbes A, Kase K, Kiisk E, et al. Diagnostic accuracy of biomarkers to detect acute mesenteric ischaemia in adult patients: a systematic review and meta-analysis. *World J Emerg Surg*. 2023;18:44.
7. Lee MH, Pickhardt PJ, Sorensen AM, Toia GV, Lubner MG, Sappenfield R, et al. Acute mesenteric ischemia: pathophysiology-based approach to imaging findings and diagnosis. *Radiographics*. 2025;45:e250012.
8. Tilsed JV, Casamassima A, Kurihara H, Mariani D, Martinez I, Pereira J, et al. ESTES guidelines: acute mesenteric ischaemia. *Eur J Trauma Emerg Surg*. 2016;42:253-70.
9. Shi Y, Zhao B, Zhou Y, Chen L, Su H, Gu J, et al. Endovascular revascularization vs open surgical revascularization as the first strategy for arterial acute mesenteric ischemia: a systematic review and meta-analysis. *J Vasc Surg*. 2024;80:1883-93.e2.
10. Harmantepe AT, Dulger UC, Gonullu E, Dikicier E, Şentürk A, Eröz E. A method for predicting mortality in acute mesenteric ischemia: machine learning. *Ulus Travma Acil Cerrahi Derg*. 2024;30:487-92.
11. Song L, Zhang X, Zhang J, Wu J, Wang J, Wang F. Deep learning-assisted diagnosis of acute mesenteric ischemia based on CT angiography images. *Front Med (Lausanne)*. 2025;12:1510357.
12. Gül Ş, Erdemir İ, Hancı V, Aydoğmuş E, Erkoç YS. How artificial intelligence can provide information about subdural hematoma: assessment of readability, reliability, and quality of ChatGPT, BARD, and Perplexity responses. *Medicine (Baltimore)*. 2024;103:e38009.
13. Hancı V, Ergün B, Gül Ş, Uzun Ö, Erdemir İ, Hancı FB. Assessment of readability, reliability, and quality of ChatGPT, BARD, Gemini, Copilot, Perplexity responses on palliative care. *Medicine (Baltimore)*. 2024;103:e39305.
14. Ömür Arça D, Erdemir İ, Kara F, Shermatov N, Odacioğlu M, İbişoğlu E, et al. Assessing the readability, reliability, and quality of artificial intelligence chatbot responses to the 100 most searched queries about cardiopulmonary resuscitation: an observational study. *Medicine (Baltimore)*. 2024;103:e38352.
15. Gianola S, Barger S, Castellini G, Cook C, Palese A, Pillastrini P, et al. Performance of ChatGPT compared to clinical practice guidelines in making informed decisions for lumbosacral radicular pain: a cross-sectional study. *J Orthop Sports Phys Ther*. 2024;54:222-8.
16. Nian PP, Saleet J, Magruder M, Wellington IJ, Choueika J, Houten JK, et al. ChatGPT as a source of patient information for lumbar spinal fusion and laminectomy: a comparative analysis against Google Web Search. *Clin Spine Surg*. 2024;37:E394-403.
17. Yıldız HA, Söğütülen E. AI chatbots as sources of STD information: a study on reliability and readability. *J Med Syst*. 2025;49:43.
18. Ozturk GA, Gurlek Tumer E, Kurt Çoban FS, Ozturk H, Gurlek U, Ozturk M, et al. AI chatbots for patient education in localized prostate cancer radiotherapy: a comparative quality and readability analysis. *Patient Educ Couns*. 2026;149:109594.
19. Dadi S, Kring T, Latz K, Cohen D, Thaller S. Evaluating the reliability and readability of AI chatbot responses for microtia patient education. *J Craniofac Surg*. 2026;37:794-7.
20. Silberg WM, Lundberg GD, Musacchio RA. Assessing, controlling, and assuring the quality of medical information on the Internet. *JAMA*. 1997;277:1244-5.
21. Charnock D, Shepperd S, Needham G, Gann R. DISCERN: an instrument for judging the quality of written consumer health information on treatment choices. *J Epidemiol Community Health*. 1999;53:105-11.
22. Bernard A, Langille M, Hughes S, Rose C, Leddin D, Veldhuyzen van Zanten S. A systematic review of patient inflammatory bowel disease information resources on the World Wide Web. *Am J Gastroenterol*. 2007;102:2070-7.
23. Moulton B, Franck LS, Brady H. Ensuring quality information for patients: development and preliminary validation of a new instrument. *Health Expect*. 2004;7:165-75.
24. CHART Collaborative; Huo B, Collins G, Chartash D, Thirunavukarasu A, Flanagan A, Iorio A, et al. Reporting guideline for chatbot health advice studies: the CHART statement. *JAMA Netw Open*. 2025;8:e2530220.
25. Flesch R. A new readability yardstick. *J Appl Psychol*. 1948;32:221-33.
26. Kincaid JP, Fishburne RP, Rogers RL, Chissom BS. Derivation of new readability formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy enlisted personnel. *Research Branch Report 8-75*. Millington (TN): Naval Technical Training Command; 1975.
27. Gunning R. The technique of clear writing. New York: McGraw-Hill; 1952.
28. McLaughlin GH. SMOG grading: a new readability formula. *J Reading*. 1969;12:639-46.
29. Senter RJ, Smith EA. Automated readability index. AMRL-TR-6620. Wright-Patterson Air Force Base (OH): Aerospace Medical Research Laboratories; 1967.
30. Coleman M, Liao TL. A computer readability formula designed for machine scoring. *J Appl Psychol*. 1975;60:283-4.
31. Weiss BD. Health literacy and patient safety: help patients understand. Manual for clinicians. 2nd ed. Chicago: American Medical Association Foundation; 2007.
32. Singh AG, Singh S, Singh PP. YouTube for information on rheumatoid arthritis - a wakeup call? *J Rheumatol*. 2012;39:899-903.
33. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J Chiropr Med*. 2016;15:155-63.
34. Nogay S, Dolanbay T, Demir B, Kartal B, Özkan A. Evaluation of the educational quality and reliability of cardiopulmonary resuscitation training videos on YouTube. *Br J Hosp Med (Lond)*. 2026;87:51256.
35. Naufal M, Kannan S, Senthilkumar S, Palaniappan T, Kumar SS, Kulkarni PP. Patient education using AI: a cross-sectional study comparing ChatGPT and Google Gemini-generated patient education brochures on various surgical management of breast cancer. *Cureus*. 2025;17:e97525.