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.

Artificial Intelligence Chatbots as Sources of Dyspepsia Information: An Evaluation of Quality and Readability

Weizhizi Li
Guangxi University of Chinese Medicine
Kunrong Liu
The First Affiliated Hospital of Guangxi University of Chinese Medicine
Xiping Tang
Guangxi Medical University Cancer Hospital
Guozhong Chen
cheninjp@163.com

The First Affiliated Hospital of Guangxi University of Chinese Medicine

Research Article

Keywords: dyspepsia, artificial intelligence, large language models, chatbots, health information, information quality, readability, patient education

Posted Date: November 4th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11037733/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

Background

Large language models (LLMs) have made artificial intelligence chatbots an increasingly common source of medical and health information for the public. Patients with dyspepsia seek information on disease definitions, symptom recognition, diagnostic evaluation, treatment choices, and subsequent management, yet the quality, source transparency, and readability of content generated by different AI chatbots have not been systematically evaluated.

Objective

To compare the quality, transparency, and readability of dyspepsia-related health information generated by four mainstream AI chatbots (ChatGPT, Gemini, Copilot, and Perplexity) and assess whether their responses meet the recommended reading level for patient education materials.

Methods

A web-based cross-sectional comparative evaluation design was used. Using the MeSH term “Dyspepsia,” Google Trends was searched worldwide from August 31, 2021, to August 31, 2026. After screening, 25 core questions were entered into the four AI chatbots using the same original English wording. Information quality, patient education value, source transparency, and overall quality were assessed with DISCERN, Ensuring Quality Information for Patients (EQIP), the JAMA Benchmark Criteria, and the Global Quality Scale (GQS). Readability was assessed with the Automated Readability Index (ARI), Flesch Reading Ease Score (FRES), Gunning Fog Index (GFI), Flesch-Kincaid Grade Level (FKGL), Coleman-Liau Index (CLI), and Simple Measure of Gobbledygook (SMOG). Quality scores were compared across models using the Kruskal-Wallis test; significant overall tests were followed by Dunn post hoc pairwise comparisons with multiple-comparison correction, and ε^2 effect sizes were reported. Readability results were compared with the sixth-grade reading standard for patient education materials. Statistical analyses were performed using Python 3.10.4.

Results

The analysis included 25 core dyspepsia-related questions. DISCERN, EQIP, JAMA, and GQS scores differed significantly among the four AI chatbots, with Perplexity and Copilot performing relatively better on some measures of information quality and source transparency. None of the models met the prespecified sixth-grade reading standard for patient education materials. ARI, GFI, FKGL, CLI, and SMOG values exceeded the recommended thresholds for all models, whereas FRES values were below the recommended threshold, indicating high reading difficulty overall.

Conclusion

Mainstream AI chatbots rapidly generate structured dyspepsia-related health information, but information quality and source transparency differ across models, and inadequate readability is common to all. AI chatbots are currently better suited as supplementary tools for patients seeking health information. Broader medical use will require better source traceability, content quality, information currency, and linguistic comprehensibility.

1. Introduction

Dyspepsia is a common gastroduodenal symptom complex characterized mainly by epigastric pain or burning, postprandial fullness, and early satiation. Routine investigations identify no organic lesion that fully explains the symptoms in most patients, many of whom are ultimately classified as having functional dyspepsia (FD) [1]. Systematic reviews estimate the global prevalence of FD in adults at approximately 6%–10%, depending on the Rome criteria applied, with substantial variation across regions and populations and a persistent public health and healthcare burden [2]. Management includes Helicobacter pylori testing, indications for endoscopy, acid-suppressive therapy, and long-term management of FD. Patients therefore need clear information on disease characteristics, symptom significance, diagnostic choices, and treatment options.

Large language models (LLMs), including ChatGPT, Gemini, Copilot, and Perplexity, now provide another route through which the public can obtain medical knowledge online [3, 4]. Compared with traditional search engines, AI chatbots can generate immediate, structured, interactive responses to natural-language questions and present complex medical concepts coherently, which may support patient education and health consultation. Their use in medicine also raises quality-control concerns, including outdated knowledge, unclear evidence sources, difficulty verifying generated content, and inconsistent performance across models [3, 5].

For patient education, medical information must be reliable, traceable, and understandable as well as sufficiently informative. Comparative studies of LLMs report differences among models in the quality of medical information, whereas poor readability appears more consistent across systems [6, 7]. Studies in rhinoplasty and refractive surgery likewise found variation in accuracy, completeness, or response appropriateness, while generated text remained above the sixth-grade reading level recommended for patient education materials [8, 9]. Organization and medically structured language alone therefore do not establish that AI-generated content is trustworthy or understandable.

Systematic multi-model evidence remains limited for dyspepsia, a common gastrointestinal symptom complex. Dyspepsia-related health information must explain disease definitions and symptom characteristics while addressing patients' practical needs for understanding causes and self-management, so information quality, source transparency, and readability warrant joint assessment. Based on real-world public search behavior, this study constructed a question set and compared the quality, transparency, and overall educational value of content generated by ChatGPT, Gemini, Copilot, and Perplexity. It addressed two questions: whether the quality of dyspepsia-related health information differs across models and whether the generated text meets the recommended reading level for patient education materials.

2. Materials and Methods

2.1 Study Design and Overall Workflow

A web-based cross-sectional comparative evaluation design was used to compare the information quality, transparency, and readability of dyspepsia-related health information provided by different artificial intelligence (AI) chatbots. Data collection was completed on August 31, 2026. All interactions were conducted in English, and the study included AI chatbot platforms and publicly available search data accessible worldwide. No patients were recruited, no clinical intervention was performed, and no personal or private information was collected.

The workflow comprised screening and construction of dyspepsia-related questions, standardized collection of AI-generated responses, storage and processing of response texts, evaluation of medical information quality, readability analysis, and statistical comparison. Real-world search behavior and a multidimensional assessment framework were combined to evaluate how current mainstream LLMs generate medical content for patients seeking health information.

2.2 Construction of Dyspepsia-Related Questions

To align the study questions with real-world patient information needs, the standard disease term “Dyspepsia” was identified using Medical Subject Headings (MeSH) and used as the search topic in Google Trends. Dyspepsia-related search terms and topics of user interest were retrieved worldwide from August 31, 2021, to August 31, 2026, to capture health information needs reflected in public search behavior.

Two evaluators with medical backgrounds independently screened the candidate items. Items were excluded if they were unrelated to dyspepsia, could not be converted into a complete health consultation question, duplicated or closely resembled another item, or did not meet the study objectives. Disagreements were resolved through discussion; if consensus was not reached, a third senior evaluator made the final decision.

After screening, 25 core questions were included for AI model evaluation. The original English wording was retained without manual simplification, rewriting, or supplementation to minimize investigator influence on model outputs. The complete list of questions is presented in Table 1.

2.3 Selection of Artificial Intelligence Models

Four representative AI chatbots widely accessible to the public were selected: ChatGPT 5.5, Gemini 3.5 Flash, Perplexity, and Copilot. Model names and version information were recorded as publicly displayed by each platform on the date of data collection.

Because LLMs are continuously updated, model performance may vary over time. The findings therefore reflect only the performance of the evaluated AI systems in generating dyspepsia-related health information during the study period.

2.4 Acquisition and Standardization of Model Responses

All core questions were entered into the four chatbots using a standardized procedure to ensure fair comparisons. Each question was submitted in a new, independent conversation window. Before data collection, browser cache, browsing history, and personalized settings that might affect model output were cleared to reduce the influence of prior interactions.

All models were queried in a single-turn format. Only the first complete response was recorded, without manual editing, deletion, or supplementation. These texts formed the raw data for the information quality and readability analyses.

2.5 Quality Assessment of AI-Generated Information

DISCERN, EQIP, GQS, and JAMA were used to assess different dimensions of AI-generated medical information quality.

DISCERN assessed the reliability of health information and the quality of treatment-related content, with attention to information sources, evidence support, descriptions of treatment options, and balance among therapeutic alternatives. Total DISCERN scores range from 16 to 80, with 63–80 classified as excellent, 51–62 as good, 39–50 as fair, 27–38 as poor, and 16–26 as very poor [10].

EQIP assessed the quality of patient health education materials, including content completeness, structural organization, and information presentation. The instrument comprises 20 items, each scored as “yes” or “no,” with the final score converted to a percentage. Scores of 76–100 indicate excellent quality, 51–75 good quality, 26–50 poor quality, and 0–25 serious quality problems [11].

GQS assessed the overall quality of AI responses, including organization, comprehensibility, and practical value for patients. The 5-point scale assigns 5 to excellent, 4 to good, 3 to fair, 2 to poor, and 1 to very poor responses [12].

The JAMA Benchmark Criteria assessed medical information transparency across four domains: authorship, attribution or references, disclosure, and currency. Each item was scored as 0 or 1, for a total score of 0 to 4; higher scores indicate

greater transparency [13].

2.6 Inter-Rater Agreement and Quality Control

Two evaluators with medical training independently rated all AI-generated responses. Before formal evaluation, approximately 10% of the questions were randomly selected for pilot assessment to harmonize scoring standards and reduce differences in interpretation.

During formal scoring, evaluators were blinded to model identity and assessed only the response text. Disagreements were resolved through discussion; unresolved cases were adjudicated by a third senior evaluator.

Inter-rater agreement was assessed to evaluate the reliability of the scoring procedure and was considered acceptable, supporting the consistency of the ratings analysed in this study.

2.7 Readability Analysis

Readability analysis assessed whether AI-generated medical information met patient reading needs. Before analysis, all responses were converted to plain text, and formatting symbols, bullet points, numbering, URLs, inline citation markers, and separate reference lists were removed. These steps standardized text format without altering the substantive content of the responses.

Six readability metrics were used: ARI, FRES, GFI, FKGL, CLI, and SMOG. The metrics were calculated with the textstat 0.7.13 package in Python 3.10.4. Based on recommendations from the American Medical Association (AMA), National Institutes of Health (NIH), and U.S. Department of Health and Human Services (HHS) for patient education materials, a sixth-grade reading level was used as the acceptable standard, defined as FRES ≥ 80 and ARI, GFI, FKGL, CLI, and SMOG values below grade 6.

2.8 Statistical Analysis

Continuous variables were summarized as medians and interquartile ranges (median [IQR]), and categorical variables as frequencies and percentages. Information quality scores were compared across AI chatbots using the Kruskal-Wallis test. Statistically significant overall comparisons were followed by Dunn post hoc pairwise comparisons with correction for multiple comparisons.

The magnitude of between-model differences was quantified with ε^2 effect sizes for the Kruskal-Wallis test. Readability indices were compared with their prespecified reading benchmarks using the Wilcoxon signed-rank test, with corresponding effect sizes reported.

All analyses were performed using Python 3.10.4 and the textstat 0.7.13 package.

3. Results

3.1 Question Screening and Dataset Construction

Google Trends yielded 50 candidate search items. Independent screening by two evaluators excluded 8 items unrelated to the study topic, 14 duplicate or highly similar items, and 3 items that did not meet the prespecified criteria, leaving 25 core dyspepsia-related questions for AI chatbot evaluation.

The questions covered definitions of dyspepsia, symptoms, causes, diagnostic evaluation, treatment, and related disease management. Common information needs included “functional dyspepsia,” “dyspepsia meaning,” “dyspepsia symptoms,” “dyspepsia treatment,” and “dyspepsia causes.” The complete list is shown in Table 1.

Table 1
Google Trends candidate search terms related to dyspepsia (August 2021–August 2026, worldwide)

Dyspepsia (August 2021–August 2026, worldwide)	
TOP search term	Relative search interest
functional dyspepsia	100
dyspepsia meaning	78
dyspepsia symptoms	66
icd dyspepsia	43
what is dyspepsia	40
icd 10 dyspepsia(removed)	35
dyspepsia icd 10(removed)	35
dyspepsia treatment	25
symptoms of dyspepsia(removed)	25
dyspepsia definition(removed)	20
gastritis(removed)	19
dyspepsia adalah(removed)	18
dyspepsia syndrome	17
functional dyspepsia symptoms	17
dyspepsia causes	17
gerd(removed)	17
syndrome dyspepsia(removed)	17
nausea	16
indigestion(removed)	16
dysphagia	15
endoscopy	13
h pylori	12
dyspepsia nhs	12
meaning of dyspepsia(removed)	12
dyspepsia medicine	11
acute dyspepsia	10
alarm symptoms dyspepsia	10
chronic dyspepsia	10
dyspepsia □□□(removed)	10
causes of dyspepsia(removed)	8

Dyspepsia (August 2021–August 2026, worldwide)	
functional dyspepsia treatment	8
kode icd dyspepsia(removed)	7
what is functional dyspepsia	7
constipation(removed)	7
ibs(removed)	7
belching	6
heartburn(removed)	6
dyspnea(removed)	6
peptic ulcer(removed)	6
epigastric pain	6
dyspepsia in hindi(removed)	6
kode icd 10 dyspepsia(removed)	6
define dyspepsia(removed)	6
dyspepsia medical term	6
satiety	6
gastroparesis(removed)	5
omeprazole	5
indigestion symptoms(removed)	5
dispepsia(removed)	5
dyspepsia icd 10 code(removed)	5

3.2 Quality Assessment of AI Chatbot Information

3.2.1 Overall Quality Across Models

The four models differed across all four information quality measures: DISCERN, EQIP, JAMA, and GQS (Table 2 and Fig. 1).

For DISCERN, median scores were 34.00 (33.00, 40.00) for ChatGPT, 33.00 (31.00, 39.00) for Gemini, 41.00 (40.00, 44.00) for Copilot, and 42.00 (42.00, 42.00) for Perplexity. The Kruskal-Wallis test showed a significant difference among models ($P < 0.001$, $\varepsilon^2=0.381$). ChatGPT and Gemini were predominantly classified as poor, whereas Copilot and Perplexity were classified as fair.

EQIP scores also differed significantly across models ($P < 0.001$, $\varepsilon^2=0.565$). Perplexity had the highest median score of 80.00 (80.00, 80.00), corresponding to excellent quality; Copilot scored 70.00 (65.00, 75.00), corresponding to good quality; ChatGPT and Gemini scored 60.00 (50.00, 65.00) and 50.00 (45.00, 55.00), respectively.

JAMA transparency scores differed significantly among models ($P < 0.001$, $\varepsilon^2=0.883$). Median scores were 0.00 for ChatGPT and Gemini and 1.00 for Copilot and Perplexity, indicating marked differences in source-transparency-related

criteria.

GQS scores also differed significantly (P = 0.012, ε²=0.110). Median scores were 4.00 for ChatGPT, Gemini, and Perplexity and 3.00 for Copilot, indicating some variation in overall organization and practical value for patients.

Table 2
Quality scores of AI chatbots

Program	ChatGPT	Gemini	Copilot	Perplexity	ε²	P
DISCERN	34.00(33.00,40.00)	33.00(31.00,39.00)	41.00(40.00,44.00)	42.00(42.00,42.00)	0.381	< 0.001
EQIP	60.00(50.00,65.00)	50.00(45.00,55.00)	70.00(65.00,75.00)	80.00(80.00,80.00)	0.565	< 0.001
JAMA	0.00(0.00,0.00)	0.00(0.00,0.00)	1.00(1.00,1.00)	1.00(1.00,1.00)	0.883	< 0.001
GQS	4.00(4.00,4.00)	4.00(3.00,4.00)	3.00(3.00,4.00)	4.00(4.00,4.00)	0.110	0.012

Note: Values are median (Q1, Q3). P values are from the Kruskal-Wallis test; ε² denotes the Kruskal-Wallis effect size.

3.2.2 Pairwise Comparisons Between Models

Dunn-Bonferroni-corrected pairwise comparisons were performed for measures with statistically significant overall differences (Table 3).

For DISCERN, ChatGPT differed significantly from Copilot (P < 0.001), as did Copilot from Gemini (P < 0.001); Copilot scored higher than both ChatGPT and Gemini. ChatGPT and Perplexity did not differ significantly (P = 0.968).

For EQIP, significant differences were observed between ChatGPT and Copilot (P = 0.002), ChatGPT and Gemini (P = 0.037), and Copilot and Gemini (P < 0.001). Copilot and Perplexity had relatively high EQIP scores.

For JAMA transparency, ChatGPT differed significantly from Copilot (P < 0.001), as did Copilot from Gemini (P < 0.001). Copilot and Perplexity did not differ significantly (P = 1.000).

For GQS, only ChatGPT and Copilot differed significantly (P = 0.009); the remaining model pairs showed no significant differences.

Between-model differences were concentrated mainly among ChatGPT, Gemini, and Copilot, whereas Copilot and Perplexity did not differ significantly on most measures.

Table 3
Pairwise comparisons between AI chatbots (P values)

Pairwise comparison	DISCERN	EQIP	JAMA	GQS
ChatGPT-Copilot	< 0.001	0.002	< 0.001	0.009
ChatGPT-Gemini	1.000	0.037	1.000	0.265
ChatGPT-Perplexity	0.968	0.568	0.292	1.000
Copilot-Gemini	< 0.001	< 0.001	< 0.001	1.000
Copilot-Perplexity	1.000	1.000	1.000	1.000
Gemini-Perplexity	0.450	0.091	0.240	1.000
Note: Values are Bonferroni-adjusted P values from Dunn's test.				

3.3 Readability of AI-Generated Text

None of the dyspepsia-related health information generated by the four AI chatbots met the prespecified sixth-grade reading standard for patient education materials (Table 4 and Fig. 2). For ChatGPT, median ARI, GFI, FKGL, CL, and SMOG values were 14.73, 15.47, 12.46, 17.10, and 12.08, respectively, with a FRES of 20.73. Corresponding values were 15.88, 16.88, 13.90, 15.69, 15.11, and 25.69 for Gemini; 13.42, 14.62, 11.45, 16.53, 11.90, and 28.05 for Copilot; and 17.20, 17.27, 14.05, 18.96, 14.91, and 19.42 for Perplexity.

For every model, median ARI, GFI, FKGL, CL, and SMOG values were above the sixth-grade threshold of 6, whereas FRES values were below 80. Copilot showed lower text complexity on ARI, GFI, FKGL, and FRES, whereas Perplexity showed higher complexity on ARI, GFI, and CL.

Perplexity combined relatively high DISCERN and EQIP scores with poor readability, indicating that the quality/reliability and comprehensibility of medical information did not track together.

Table 4
Readability scores of AI chatbots

Program	ChatGPT	Gemini	Copilot	Perplexity	6th-grade threshold
ARI	14.73(13.84,15.79)	15.88(14.67,17.65)	13.42(12.23,14.09)	17.20(17.20,17.20)	< 6
GFI	15.47(13.91,16.34)	16.88(15.99,18.39)	14.62(12.98,15.12)	17.27(17.27,17.27)	< 6
FKGL	12.46(11.79,13.31)	13.90(12.99,15.06)	11.45(10.32,12.80)	14.05(14.05,14.05)	< 6
CL	17.10(15.84,17.91)	15.69(14.85,17.12)	16.53(14.74,17.81)	18.96(18.96,18.96)	< 6
SMOG	12.08(11.87,12.79)	15.11(14.31,16.13)	11.90(11.37,12.49)	14.91(14.91,14.91)	< 6
FRES	20.73(18.24,26.00)	25.69(20.21,30.78)	28.05(19.30,34.65)	19.42(19.42,19.42)	>=80
Note: Values are median (Q1, Q3). Sixth-grade threshold: FRES ≥ 80; all other grade-level indices < 6.					

3.4 Question-Level Distributions

The quality and readability heatmaps show score distributions across questions and models (Fig. 3). Within each model, quality and readability varied across questions, suggesting that response performance may depend on question content.

Because the available statistical tables did not provide formal measures of question-level variability, the heatmaps were interpreted descriptively, without inference about the statistical significance of within-model variation.

4. Discussion

4.1 Main Findings

The four AI chatbots differed in information quality and source transparency, yet all generated text above the sixth-grade reading level. DISCERN, EQIP, JAMA, and GQS scores differed significantly among models, and JAMA had the largest effect size, suggesting that source transparency may be a prominent dimension of between-model variation. Perplexity had relatively high median DISCERN and EQIP scores, although its absolute quality remained below an ideal level. All six readability metrics indicated that text from all four models was substantially above the sixth-grade reading level, making poor readability a common limitation.

4.2 Comparison With Previous Studies

The results align with recent studies of AI-generated medical information and extend the evidence to dyspepsia, a common gastrointestinal disorder.

For information quality, Chandra et al. compared ChatGPT-generated gastrointestinal patient education materials with official materials from the American Gastroenterological Association (AGA) and found no significant difference in DISCERN scores (63.5 ± 5.7 vs 64.3 ± 5.4) [14]. This result suggests that AI-generated content may approach the quality of materials produced by a professional society. By contrast, DISCERN scores for the four models in our study (median range, 33–42) were substantially lower than the ChatGPT score of 63.5 reported by Chandra et al.; disease area, question type, or model version may account for part of this difference. Çetinkaya et al. evaluated eight chatbots for brucellosis-related health information and found significant differences in clinical accuracy, source transparency, and reliability, with DeepSeek performing best across several measures [15].

Poor readability also matches previous reports. Chandra et al. found that ChatGPT-generated gastrointestinal educational materials had a significantly higher reading grade level than AGA materials (10.33 ± 1.5 vs 8.72 ± 1.7) [14]. A contemporaneous commentary noted that most U.S. adults read at or below the sixth-grade level and that the greater reading difficulty of ChatGPT-generated materials could worsen existing inequities in health literacy and access [16]. In irritable bowel syndrome, responses from ChatGPT-4 and Gemini-1 were both rated at a college-level reading difficulty, with FRES values of 32.33 and 35.83, respectively [17]. A study of patient education information on glioma similarly found median Flesch-Kincaid grade levels of 8.47 to 13.43 across six chatbots; no platform consistently achieved the sixth-grade target [18].

Perplexity paired relatively high DISCERN and EQIP scores with high text complexity, including ARI 17.20 and GFI 17.27. This pattern is consistent with Çetinkaya et al. [15], suggesting that higher technical quality does not necessarily translate into greater user trust or readability. A systematic review and exploratory meta-analysis of AI-generated spine surgery patient education materials termed this pattern a “readability paradox”: quality assessments may be acceptable even when the text remains systematically difficult to read [19].

4.3 Interpretation of Quality Differences Between Models

The observed heterogeneity in medical information quality reflects differences among current LLMs in system architecture, information-source integration, and optimization strategies.

Source transparency showed the largest between-model effect. JAMA had the largest effect size (ε^2=0.883); median scores were 1.00 for both Copilot and Perplexity and 0.00 for both ChatGPT and Gemini. This difference may relate to

information retrieval mechanisms. Perplexity and Copilot can integrate real-time web retrieval and external knowledge sources, making source and citation cues easier to include, whereas models relying mainly on pretrained parameters, such as base versions of ChatGPT, may omit sources or provide unclear evidence attribution. Research on medical LLMs likewise identifies opacity in training data sources, knowledge updating pathways, and output verifiability as a limitation to medical use despite strong medical knowledge integration [3, 20].

Information integration and presentation may also contribute to performance differences. Copilot had higher DISCERN (41.00) and EQIP (70.00) scores than ChatGPT (34.00 and 60.00) and Gemini (33.00 and 50.00), which may relate to stronger information retrieval and integration capabilities. In a comparative study of responses to neurological questions, Copilot likewise outperformed other models [21].

Model rankings are context-dependent. Different LLMs do not show a stable ranking across medical question-answering tasks, and previous studies suggest that performance may depend on pretraining data, instruction tuning, reinforcement learning from human feedback, and external information access. The relative advantages observed here should therefore be interpreted within this research context and evaluation framework rather than as evidence of universally superior medical accuracy or clinical reliability.

4.4 Causes and Consequences of Inadequate Readability

All four AI chatbots generated dyspepsia-related health information that failed to meet the recommended reading level for patient education materials.

Medical information is inherently specialized. Dyspepsia encompasses symptom recognition, disease classification (functional vs organic dyspepsia), Helicobacter pylori testing, indications for endoscopy, and treatment selection, all of which require medical terminology and relatively complex concepts. Even with natural-language generation, some terminology and logical relationships may remain beyond the comprehension of average patients. LLMs are also generally optimized to generate complete, logically structured responses. Explanations of disease mechanisms and diagnostic or treatment pathways may therefore increase information density and linguistic complexity rather than adapt to patients' reading ability. Some research suggests that models optimized for factual precision generate denser language, whereas more conversational models may be more fluent but sacrifice depth; this tension between "accuracy" and "readability" is an inherent challenge in AI-generated medical information.

Poor readability may limit the practical value of AI-generated health information. Information that patients cannot understand is unlikely to translate into effective health behavior even when medically accurate. High reading difficulty may impair understanding of disease, treatment choices, and indications for seeking care, reducing the effectiveness of AI as an adjunctive health education tool. In gastroenterology, a study in inflammatory bowel disease found that although ChatGPT responses were comparable with gastroenterologists in accuracy and completeness, patients rated treatment-related responses significantly lower than those provided by experts, pointing to readability and patient trust as relevant dimensions of AI-generated medical information [22].

4.5 Practical Implications

The findings have practical implications for patients, clinicians, and AI developers.

For patients, AI chatbots may help provide basic disease information, organize health-related questions, and support patient-clinician communication, but they should not be regarded as independent sources of diagnosis or treatment. Because medical information transparency and readability remain limited, decisions involving diagnostic judgments, medication choices, testing needs, or management of changing symptoms should still be informed by authoritative medical sources and professional consultation.

Patients increasingly use AI to obtain health information. Clinicians may therefore ask whether patients have used AI tools for disease-related information and address any misunderstandings. Professionally reviewed, easy-to-understand educational materials tailored to patients' needs may also improve comprehension and use of disease-management information.

For AI developers, optimization of medical LLMs should address transparency and traceability alongside content generation. Model outputs should, whenever possible, identify information sources, referenced guideline versions, and dates of update. Plain-language constraints and health-literacy adaptation strategies should reduce reading barriers and support individualized expression based on users' health literacy, with the aim of improving safety and practical value in patient education settings.

4.6 Strengths and Limitations

The study design has several strengths. Using Google Trends to identify dyspepsia-related search terms of actual public interest, rather than investigator-designed questions alone, better reflects real-world information-seeking behavior. The assessment combined DISCERN, EQIP, JAMA, and GQS with six readability metrics to evaluate information quality, transparency, overall practical value, and text complexity. The statistical analysis included nonparametric overall comparisons, effect sizes, and post hoc pairwise comparisons, reducing reliance on a single mean score to rank models.

The study also has limitations. Only four AI chatbots were evaluated, so the findings cannot be generalized to all LLMs. The 25 questions selected through Google Trends cannot capture all individualized patient needs, and Google Trends provides no demographic information on age, sex, educational level, or health literacy. Because LLMs are continuously updated, the findings reflect model performance only at the specific time point of August 31, 2026; future versions may perform differently, so longitudinal and periodic reassessment is warranted. DISCERN, EQIP, JAMA, GQS, and traditional readability metrics were developed primarily for static written materials and may have limited applicability to dynamic, conversational AI content. Readability formulas rely mainly on sentence length, word length, syllables, or characters and cannot substitute directly for actual patient comprehension. Future research should involve target patient populations and assess comprehension, trust, information retention, and effects on decision-making.

5. Conclusion

Information quality and source transparency differed significantly among ChatGPT, Gemini, Copilot, and Perplexity, although all four rapidly generated structured responses to common public questions about dyspepsia. Perplexity performed relatively well on DISCERN and EQIP, and Copilot showed advantages on several measures; a higher relative ranking, however, does not indicate ideal medical information quality. None of the four AI chatbots produced dyspepsia information that met the sixth-grade reading level recommended for patient education materials, so excessive linguistic complexity remains a common limitation. AI chatbots are currently better suited as supplementary tools for patient health education than as substitutes for professional medical consultation. Model development should improve medical content quality, source traceability, information currency, and linguistic comprehensibility, and should explore individualized communication based on users' health literacy.

Abbreviations

The following abbreviations are used in this manuscript:

AI	artificial intelligence
AMA	American Medical Association

ARI Automated Readability Index

CLI Coleman-Liau Index

EQIP Ensuring Quality Information for Patients

FD functional dyspepsia

FKGL Flesch-Kincaid Grade Level

FRES Flesch Reading Ease Score

GFI Gunning Fog Index

GQS Global Quality Scale

HHS U.S. Department of Health and Human Services

IQR interquartile range

JAMA Journal of the American Medical Association (JAMA Benchmark Criteria)

LLMs large language models

MeSH Medical Subject Headings

NIH National Institutes of Health

SMOG Simple Measure of Gobbledygook

Declarations

Ethics approval

Ethics approval declaration: not applicable. This study did not involve human participants, human data, or human tissue. All material evaluated consisted of text generated by publicly available artificial intelligence chatbots in response to investigator-defined questions, together with publicly accessible aggregate search-volume data from Google Trends. The two independent evaluators and the third senior evaluator acted as investigators performing blinded scoring rather than as research subjects. Because no research involving human subjects was conducted, approval was not sought from any ethics committee or institutional review board (IRB); accordingly, no approval committee or IRB name and no approval number can be provided. The study was conducted in accordance with the principles of the Declaration of Helsinki.

Consent to participate

Consent to Participate declaration: not applicable. Every human participant in a study involving human subjects is required to provide their consent; however, this study enrolled no human participants. No patients or healthy volunteers were recruited, surveyed, sampled, or interviewed, no intervention was performed, and no identifiable personal or clinical data were collected, accessed, or analysed. All AI-generated responses were obtained from publicly available chatbot interfaces, and all Google Trends data were publicly accessible and non-identifiable aggregate data. Individual consent to participate was therefore not required and was not obtained.

Human Ethics and Consent to Participate declarations

Human Ethics and Consent to Participate declarations: not applicable. Neither human ethics approval nor consent to participate was required, for the reasons stated in the "Ethics approval" and "Consent to participate" declarations above.

Consent for publication

Not applicable. This manuscript contains no individual person's data in any form, including individual clinical details, images, or videos.

Author contributions

WL: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Visualization, Writing – original draft. XT: Investigation, Validation, Writing – review & editing. KL: Investigation, Validation, Writing – review & editing. GC: Conceptualization, Supervision, Writing – review & editing. All authors contributed to the article and approved the submitted version.

Funding

The author(s) declared that financial support was received for this work and/or its publication. This work was supported by the National Natural Science Foundation of China (Nos. 82560872 and 82661045).

Data availability statement

The datasets generated and analysed during the current study are not publicly deposited, as they consist of verbatim outputs generated by commercial artificial intelligence chatbot platforms; however, they are available from the corresponding author on reasonable request. All data supporting the findings of this study are presented within the article and its appendix

Acknowledgments

The authors thank the two independent evaluators and the third senior evaluator for their assistance with the blinded scoring of the AI-generated responses.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Black CJ, Paine PA, Agrawal A, Aziz I, Eugenicos MP, Houghton LA, Hungin P, Overshott R, Vasant DH, Rudd S, et al. British Society of Gastroenterology guidelines on the management of functional dyspepsia. Gut. 2022;71:1697–723. https://doi.org/10.1136/gutjnl-2022-327737.
2. Lee K, Kwon CI, Yeniova AÖ, Koyanagi A, Jacob L, Smith L, Lee SW, Rahmati M, Shin JY, Shin JI, et al. Global prevalence of functional dyspepsia according to Rome criteria, 1990–2020: A systematic review and meta-analysis. Sci Rep. 2024;14:4172. https://doi.org/10.1038/s41598-024-54716-3.
3. Giuffrè M, Kresevic S, You K, Dupont J, Huebner J, Grimshaw AA, Shung DL. Systematic review: The use of large language models as medical chatbots in digestive diseases. Aliment Pharmacol Ther. 2024;60:144–66. https://doi.org/10.1111/apt.18058.
4. Klang E, Sourosh A, Nadkarni GN, Sharif K, Lahat A. Evaluating the role of ChatGPT in gastroenterology: A comprehensive systematic review of applications, benefits, and limitations. Ther Adv Gastroenterol.

2023;16:17562848231218618. https://doi.org/10.1177/17562848231218618.

5. Toiv A, Saleh Z, Ishak A, Alsheik E, Venkat D, Nandi N, Zuchelli TE. Digesting digital health: A study of appropriateness and readability of ChatGPT-generated gastroenterological information. Clin Transl Gastroenterol. 2024;15:e00765. https://doi.org/10.14309/ctg.0000000000000765.
6. Seth I, Lim B, Xie Y, Cevik J, Rozen WM, Ross RJ, Lee M. Comparing the efficacy of large language models ChatGPT, BARD, and Bing AI in providing information on rhinoplasty: An observational study. Aesthet Surg J Open Forum. 2023;5:ojad084. https://doi.org/10.1093/asjof/ojad084.
7. Meyer MKR, Kandathil CK, Davis SJ, Durairaj KK, Patel PN, Pepper JP, Spataro EA, Most SP. Evaluation of rhinoplasty information from ChatGPT, Gemini, and Claude for readability and accuracy. Aesthetic Plast Surg. 2025;49:1868–73. https://doi.org/10.1007/s00266-024-04343-0.
8. Nguyen AT, Li RA, Galiano RD. Quality and readability of online rhinoplasty information: A systematic review and meta-analysis. Ann Plast Surg. 2025;95:478–84. https://doi.org/10.1097/SAP.0000000000004441.
9. Aydın FO, Aksoy BK, Ceylan A, Akbaş YB, Ermiş S, Kepez Yıldız B, Yıldırım Y. Readability and appropriateness of responses generated by ChatGPT 3.5, ChatGPT 4.0, Gemini, and Microsoft Copilot for FAQs in refractive surgery. Turk J Ophthalmol. 2024;54:313–7. https://doi.org/10.4274/tjo.galenos.2024.28234.
10. Charnock D, Shepperd S, Needham G, Gann RDISCERN. An instrument for judging the quality of written consumer health information on treatment choices. J Epidemiol Community Health. 1999;53:105–11. https://doi.org/10.1136/jech.53.2.105.
11. 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;7:165–75. https://doi.org/10.1111/j.1369-7625.2004.00273.x.
12. Bernard A, Langille M, Hughes S, Rose C, Leddin D, Veldhuyzen, van Zanten. A systematic review of patient inflammatory bowel disease information resources on the World Wide Web. Am J Gastroenterol. 2007;102:2070–7. https://doi.org/10.1111/j.1572-0241.2007.01325.x.
13. Silberg WM, Lundberg GD, Musacchio RA. Assessing, controlling, and assuring the quality of medical information on the Internet: Caveant lector et viewor—Let the reader and viewer beware. JAMA. 1997;277:1244–5. https://doi.org/10.1001/jama.1997.03540390074039.
14. Chandra S, Kumar V, Sapkota A, Kwei-Nsoro R, Almoghrabi A. Readability, quality, understandability, and actionability of ChatGPT generated GI patient education versus AGA Patient Center. Dig Dis Sci. 2026. https://doi.org/10.1007/s10620-026-10087-5.
15. İzci Çetinkaya F, Mirza A, Ekici H, Eryılmaz Eren E, Ture Z. Evaluation of artificial intelligence chatbots in providing brucellosis-related health information: A multidimensional quality assessment. Zoonoses Public Health. 2026;73:380–92. https://doi.org/10.1111/zph.70059.
16. Flanagan RF. Concise commentary: AI-generated GI patient education-high quality, yet readability may limit utility. Dig Dis Sci. 2026. https://doi.org/10.1007/s10620-026-10152-z.
17. Raghavendran AK, Musunuri B, Rajpurohit S, Ganesh Pai C, Shetty S, Kumari P, Shetty R, Shetty A, Bhat G. Evaluation of artificial intelligence-based patient education models for irritable bowel syndrome. Indian J Gastroenterol. 2026;45:918–25. https://doi.org/10.1007/s12664-025-01872-7.
18. Shukla IY, Sun MZ. Online and ChatGPT-generated patient education materials regarding brain tumor prognosis fail to meet readability standards. J Clin Neurosci. 2025;138:111410. https://doi.org/10.1016/j.jocn.2025.111410.
19. Alaika AA, Novitasari RK, Sugiharto. The readability paradox and illusion of quality in AI-generated spine surgery patient education: A systematic review and exploratory meta-analysis. Health Policy Technol. 2026;15:101244. https://doi.org/10.1016/j.hlpt.2026.101244.

20. Liu S, McCoy AB, Wright A. Improving large language model applications in biomedicine with retrieval-augmented generation: A systematic review, meta-analysis, and clinical development guidelines. J Am Med Inf Assoc. 2025;32:605–15. https://doi.org/10.1093/jamia/ocaf008.
21. Shin D, Park H, Shaffrey I, Yacoubian V, Taka TM, Dye J, Danisa O. Artificial intelligence versus clinical judgement: How accurately do generative models reflect CNS guidelines for Chiari malformation? Clin Neurol Neurosurg. 2025;248:108662. https://doi.org/10.1016/j.clineuro.2024.108662.
22. He Z, Liu Y, Wang Z, Zhou X, Fan X, He M, Wang C, Fu H, Wang Z, Bai Y. Revolutionizing inflammatory bowel disease healthcare communication: A head-to-head comparison of gastroenterologist and ChatGPT responses. Gastroenterol Rep. 2025;13:goaf087. https://doi.org/10.1093/gastro/goaf087.

Figures

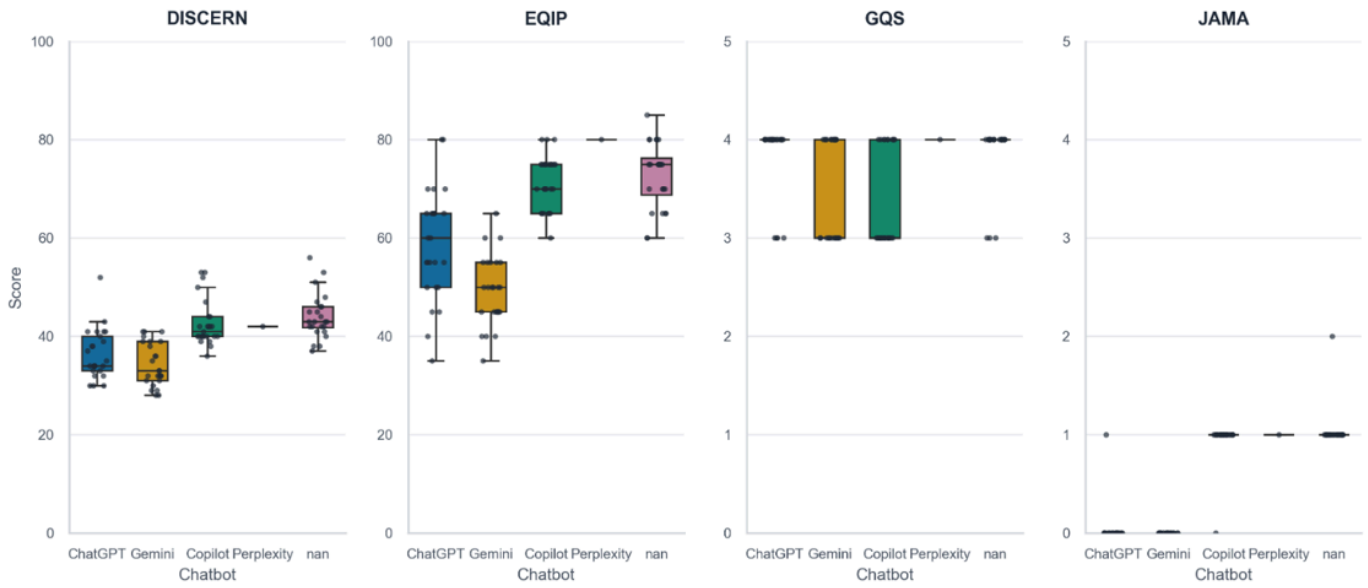


Figure 1

Distribution of DISCERN, EQIP, GQS, and JAMA scores across the four AI chatbots.

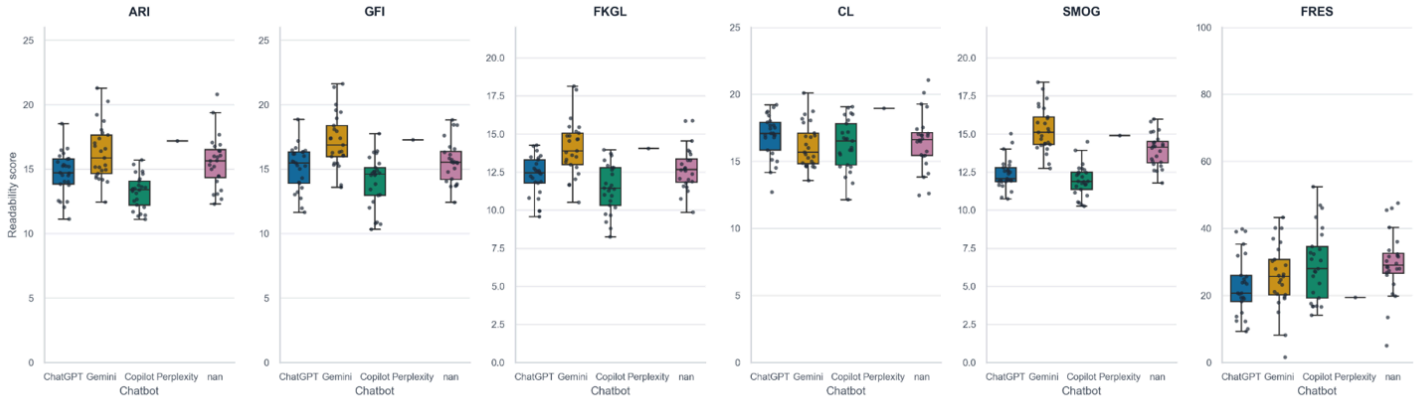


Figure 2

Distribution of six readability metrics across the four AI chatbots.

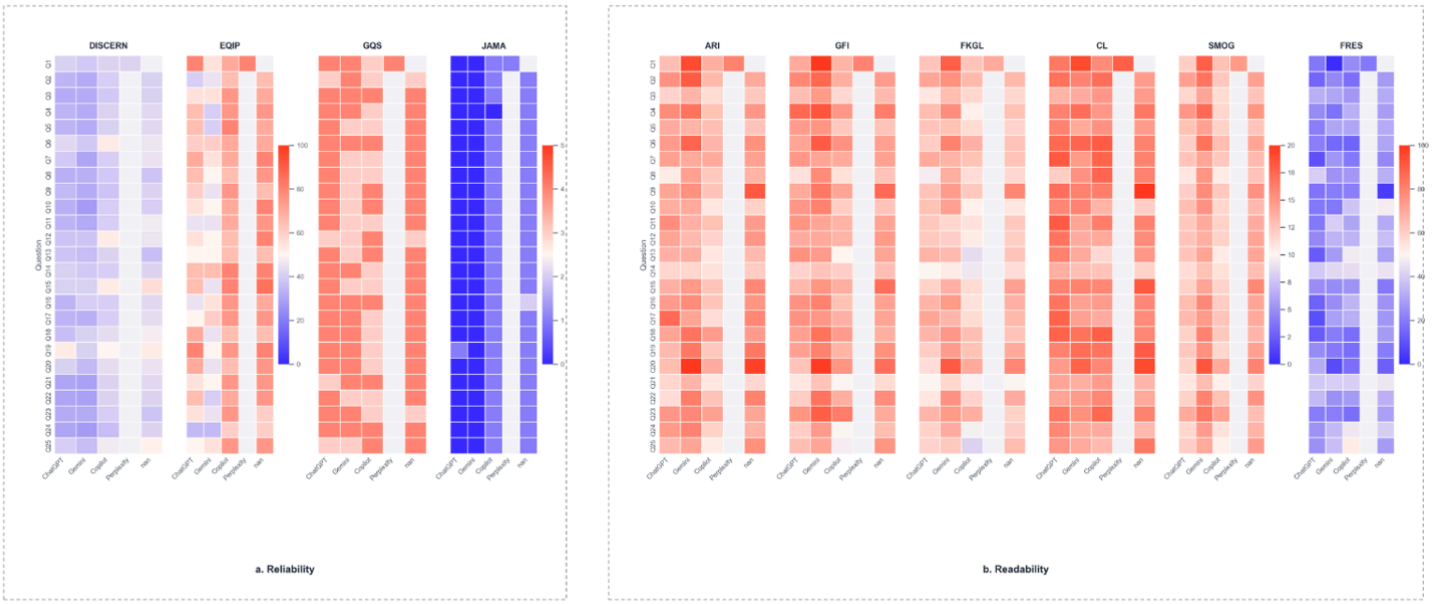


Figure 3

Heatmaps of quality and readability across different questions and models.

Supplementary Files

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

- AppendixA.docx