Original Paper

Safety, Accuracy, Empathy, Information Quality, and Readability of Chronic Obstructive Pulmonary Disease Health Advice Generated by 6 Large Language Models: Cross-Sectional Comparative Evaluation Study

Liping Fan[1], BM; Ru Liu[1]; Zhishui Wu[2], MN; Jingjing Li[1]; Jiao Lei[1]; Qingqing Zhang[1]; Xinghua Li[1], MMed; Yanming Yang[3], MNS

[1]Department of Respiratory and Critical Care Medicine, The First Affiliated Hospital of Henan University of Chinese Medicine, Zhengzhou, Henan, China

[2]Department of Spine Surgery, The First People's Hospital of Changzhou, Changzhou, Jiangsu, China

[3]Department of Nursing, The First Affiliated Hospital of Henan University of Chinese Medicine, Zhengzhou, Henan, China

Corresponding Author:
Yanming Yang, MNS
Department of Nursing
The First Affiliated Hospital of Henan University of Chinese Medicine
No. 19 Renmin Road, Jinshui District
Zhengzhou, Henan, 450000
China
Phone: 86 15936218289
Email: yangyanming-228@126.com

Abstract

Background: Large language models (LLMs) are increasingly used for health information, but chronic obstructive pulmonary disease (COPD)–related questions often involve safety-sensitive decisions, including treatment escalation, oxygen therapy, and exacerbation management. Because generative outputs may vary across repeated prompts, response stability is also relevant to comparative evaluation.

Objective: This study aims to compare 6 public-facing LLM chatbots in answering COPD-related public health questions across safety, accuracy, empathy, information quality, source transparency, readability, and 2-round response stability.

Methods: We developed a 55-question English COPD question bank from an initial 108-question pool using web-based retrieval, Google Trends, PubMed-informed topic review, patient education resources, and expert screening. Six public-facing chatbot products were evaluated in a primary first round and a post hoc independent second round, added after first-round data collection, using the same 55 questions. Visible model/mode labels are provided in the supplementary materials. Five blinded raters assessed manually scored outcomes against a Global Initiative for Chronic Obstructive Lung Disease (GOLD) 2026-centered reference standard, and 6 readability indices were calculated. The primary cross-model analysis used the first round; the second round was analyzed as a matched retest sensitivity analysis. Stability was assessed using exact agreement and Cohen κ for safety, quadratic-weighted κ for ordinal ratings, and absolute-agreement single-measure intraclass correlation coefficients (ICCs) for continuous scores and readability indices; 95% CIs used question-level bootstrap resampling.

Results: The first round yielded 330 responses. Unsafe responses occurred in every model: ChatGPT, Gemini, Copilot, and DeepSeek each generated 4 unsafe responses (7.3%); Qwen generated 5 (9.1%); and Doubao generated 6 (10.9%). The overall safety comparison was not statistically significant (Cochran Q=1.909, df=5; *P*=.86); this finding should not be interpreted as evidence of equivalent safety. Accuracy, empathy, information-quality measures, source-transparency scores, and readability indices differed significantly across models. Across 330 matched question-model pairs, safety classifications agreed in 97.3% of pairs (κ=0.821, 95% CI 0.543-0.979); accuracy and empathy showed weighted κ values of 0.872 and 0.856, respectively; DISCERN (an instrument for judging the quality of written consumer health information on treatment choices) and adapted Ensuring Quality Information for Patients (EQIP) showed ICCs of 0.954 and 0.925. Most readability indices also showed high overall 2-round stability, although model-specific variability remained. Nine safety classifications changed across the 2 rounds (5 from safe to unsafe and 4 from unsafe to safe).

XSL•FO
RenderX

Conclusions: In this standardized single-turn COPD question set, model performance differed across several evaluated outcomes, whereas the first-round safety test neither established a between-model difference nor demonstrated equivalence. All 6 models produced potentially unsafe advice, and 9 safety classifications changed between rounds conducted several months apart despite high pooled agreement. These findings support cautious use of LLMs for supplementary COPD education, with clinician oversight for individualized treatment decisions and ongoing reassessment as public-facing systems change.

(JMIR Med Inform 2026;14:e104836) doi: 10.2196/104836

KEYWORDS

CHART; Chatbot Assessment Reporting Tool; chatbot; chronic obstructive pulmonary disease; COPD; health information quality; large language model; LLM; patient education; readability; safety

Introduction

Chronic obstructive pulmonary disease (COPD) is a major cause of chronic symptoms, disability, hospitalization, and premature death worldwide [1-5]. The global prevalence of COPD remains high, with a disproportionate burden in low- and middle-income countries, and its future health burden is expected to remain substantial [2,4,5]. Contemporary guidelines emphasize that COPD is preventable and treatable, but appropriate management requires accurate diagnosis, risk-factor reduction, individualized pharmacological treatment, pulmonary rehabilitation, vaccination, regular follow-up, and timely management of exacerbations [1,3,6,7].

The public increasingly turns to online platforms and generative AI tools for health information [8-10]. Large language models (LLMs) can provide rapid, fluent, and seemingly personalized responses, which may make them attractive for patients and family caregivers seeking information about COPD definitions, risk factors, symptoms, testing, inhaled therapy, oxygen use, pulmonary rehabilitation, and exacerbation management [11-21]. However, fluency does not guarantee clinical accuracy, safety, transparency, empathy, or readability [22-24].

COPD public counseling poses a particular challenge for generative AI evaluation. Many apparently simple COPD questions are condition-dependent and clinically nuanced. Whether a patient needs inhaled corticosteroids, triple therapy, antibiotics during an exacerbation, long-term oxygen therapy, noninvasive ventilation, or urgent medical review depends on symptom severity, exacerbation history, spirometry, oxygenation, blood eosinophils, comorbidities, prior treatment response, and individualized clinician assessment [1,3,6,7]. Thus, factual plausibility alone is an insufficient marker of response quality. Recent COPD-specific evaluations have similarly shown that LLM performance varies across systems and across outcome domains, including accuracy, comprehensiveness, information quality, and readability [25-28].

In addition, COPD self-management is strongly influenced by health literacy, and patient education materials should be understandable, actionable, and aligned with evidence-based recommendations [29-32]. A response may be technically accurate yet clinically unsafe if it encourages self-directed treatment escalation, or unusable if it exceeds the reading level appropriate for public education [29-32]. These issues are highly relevant in COPD, where misunderstanding can affect inhaler use, recognition of acute exacerbations, adherence, smoking cessation, vaccination, and timely care seeking.

Previous studies evaluating medical chatbots and LLMs have reported encouraging performance in some health-related question-answering tasks [11-14,16-19]. However, systematic reviews and broader clinical evaluations have also identified important limitations in accuracy, empathy, source attribution, consistency, and domain-specific safety [20-24]. Moreover, many existing studies focus on general medical questions or single quality dimensions, whereas relatively few have specifically examined COPD-related public health consultation using a multidimensional framework that simultaneously captures safety, accuracy, empathy, information quality, transparency, and readability. Validated or widely used appraisal tools for consumer health information, including DISCERN (an instrument for judging the quality of written consumer health information on treatment choices), Ensuring Quality Information for Patients (EQIP), Journal of the American Medical Association (JAMA) benchmark criteria, and Global Quality Score (GQS), provide a methodological basis for such multidimensional evaluation [32-35]. The CHART (Chatbot Assessment Reporting Tool) framework further emphasizes the need for transparent reporting of chatbot health-advice studies [36].

This study therefore aimed to compare 6 widely used public-facing LLM chatbots in answering COPD-related public health consultation questions across safety, accuracy, empathy, information quality, source transparency, and readability. In response to the stochastic nature of generative systems, we additionally conducted a complete second query round and evaluated matched 2-round response stability. The target audience was the general public seeking COPD information; the intervention was each chatbot response; the comparators were the other LLMs; and outcomes included guideline-aligned safety and accuracy, empathic communication, information quality, source transparency, readability, and response stability.

Methods

Study Design and Reporting Framework

This was a CHART-guided cross-sectional comparative evaluation of 6 generative AI-driven chatbots for public-facing COPD health advice [23,36], supplemented by a post hoc, matched second-round retest stability analysis added after completion of first-round data collection. The primary unit for cross-model comparisons was the response generated by one

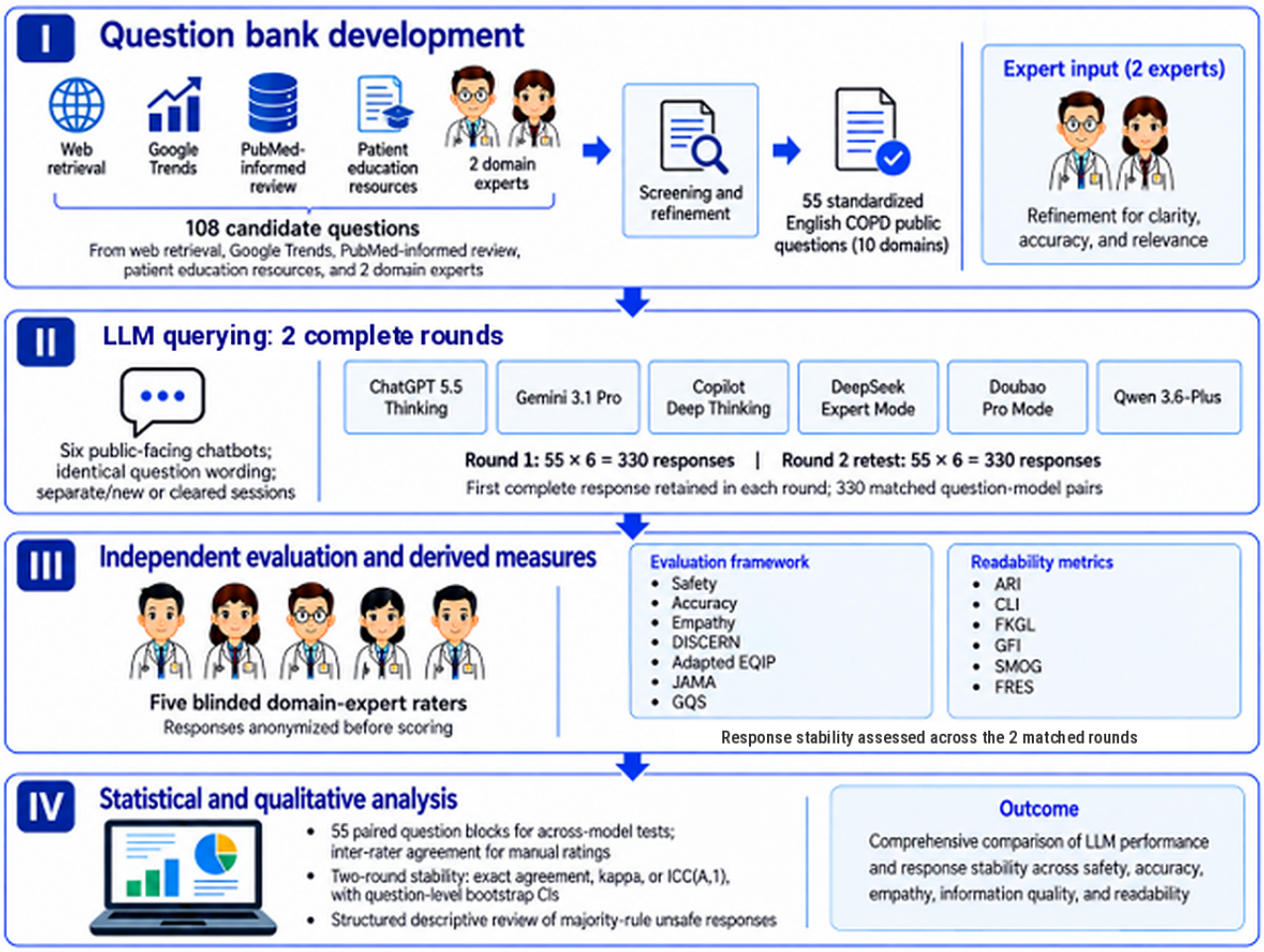
XSL•FO
RenderX

model to one standardized COPD-related public question in the first round, with the question serving as the paired block across models. The second round created 330 matched question-model pairs for stability analysis. Quantitative ratings were complemented by a structured descriptive review of responses classified as unsafe.

A concise summary of the study protocol is provided in Multimedia Appendix 1. The completed CHART checklist is provided in Multimedia Appendix 2.

The study did not recruit patients, collect patient data, use biological specimens, or intervene in clinical care. It used study-generated public questions and chatbot-generated text outputs. Because the evaluated systems are frequently updated, all findings should be interpreted as time-specific observations of public-facing interfaces rather than permanent model characteristics. The overall study workflow is summarized in Figure 1, which reflects the 55-question development process, 2 complete query rounds across 6 models, 5-rater evaluation framework, readability assessment, matched stability analysis, and statistical/qualitative analysis plan.

Figure 1. Study workflow for question-bank development, large language model (LLM) querying, independent evaluation, and statistical analysis. ARI: Automated Readability Index; CLI: Coleman-Liau Index; COPD: chronic obstructive pulmonary disease; EQIP: Ensuring Quality Information for Patients; FKGL: Flesch-Kincaid Grade Level; FRES: Flesch Reading Ease Score; GFI: Gunning Fog Index; GQS: Global Quality Score; ICC: intraclass correlation coefficient; JAMA: Journal of the American Medical Association; SMOG: Simple Measure of Gobbledygook.



Question Identification, Prompt Development, and Sample-Size Rationale

The question bank was developed to reflect public-facing COPD concerns rather than specialist-only clinical scenarios. Candidate questions were derived from broad web-based retrieval, Google Trends, PubMed-informed topic review, patient education resources, and suggestions from 2 domain experts. The retrieval range covered April 2025 to April 2026, and the search was conducted in May 2026. The initial pool contained 108 dimension-classified candidate questions.

Questions were screened for topic relevance, duplication, public-facing wording, clinical interpretability, and breadth of COPD domain coverage. The final 55-question set covered 10 dimensions: disease basics and definition, risk factors and causes, symptoms and clinical presentation, diagnosis and testing, treatment—medicines, treatment—nonmedicine, acute exacerbations, prevention and risk reduction, monitoring and follow-up, and complications and comorbidities. The final number of 55 questions was determined pragmatically based on domain coverage, expert consensus, removal of duplicate items, and feasibility for full blinded scoring across all models. The complete 55-question COPD question bank is provided in Multimedia Appendix 3. The complete 108-to-55 inclusion and exclusion record is provided in Multimedia Appendix 4.

XSL•FO
RenderX

The study prompt was the question itself, entered without additional leading instructions, role assignment, or answer-format constraints, to approximate how a member of the public might use a chatbot. No patient or public representative directly participated in prompt development.

Chatbots, Model Identifiers, Access Route, and Query Standardization

Six LLM chatbots were evaluated. All were accessed through their user-facing chatbot interfaces. The reported model identifiers refer to the product names and selectable modes visible in the user-facing interfaces at the time of querying. The evaluated public-facing chatbot interfaces were treated as externally hosted systems whose exact backend configurations, model weights, deployment parameters, and update histories were not accessible to the investigators. This description refers to the evaluated user-facing interfaces and does not necessarily define the licensing status of related model families. Official release dates or last-update dates were not available from the user-facing interfaces. The underlying base-model, tuned-model, or fine-tuned-model status of each public-facing chatbot was not independently verifiable through the user interface. No model was trained, tuned, or fine-tuned by the investigators. Chatbot versions, access type, source status, and query dates are provided in Multimedia Appendix 5. All chatbot queries were conducted by the investigators from Zhengzhou, Henan Province, China. The query dates were as follows: Copilot Deep Thinking and Doubao Pro Mode were queried on May 12, 2026; DeepSeek Expert Mode was queried on May 13, 2026; ChatGPT 5.5 Thinking and Qwen 3.6-Plus were queried on May 15, 2026; and Gemini 3.1 Pro was queried on May 19, 2026. The reported query location refers to the investigator-side access location rather than the server location of the chatbot providers.

In the first round, each question was submitted once to each chatbot in a separate new or cleared chat session to reduce contextual contamination. The same question wording and sequence were used across models, only the first complete response was retained, and no follow-up prompting, regeneration, or manual correction was performed. Prompt/question wording was reviewed by 2 domain experts. First-round queries were conducted across May 12 to 19, 2026, rather than in a single synchronized time window. The second-round retest was conducted on September 1 and 2, 2026, using the same 55 standardized questions across the same 6 public-facing chatbot products. This retest was added after completion of first-round data collection and was not specified in the original study protocol; it was therefore analyzed as an exploratory retest sensitivity analysis rather than as a prespecified primary analysis. The visible modes or version labels available during the second round are reported separately in Multimedia Appendix 6. As in round 1, the first complete response for each question-model pair was retained in a separate new or cleared session. The resulting retest dataset contained 330 matched question-model responses. Because public-facing systems can be updated without notice and exact backend parameters are not exposed, the 2-round analysis is interpreted as a pragmatic stability assessment of the evaluated interfaces rather than a complete characterization of within-model stochasticity. The complete list of study prompts is provided in Multimedia Appendix 3; first- and second-round chatbot details are provided in Multimedia Appendices 5 and 6.

The compiled question-and-answer records for the COPD LLM responses are provided in Multimedia Appendix 7.

Reference Standard and Performance Evaluation

The primary reference standard was a Global Initiative for Chronic Obstructive Lung Disease (GOLD) 2026-centered COPD question-answer mapping table. Each of the final 55 questions was mapped to a guideline-consistent answer and corresponding source references. Responses were judged against this reference framework as well as general respiratory clinical principles. The complete COPD question-answer mapping table is provided in Multimedia Appendix 8.

Five blinded reviewers with domain-specific expertise independently evaluated the anonymized LLM responses. The reviewers were from 3 different tertiary hospitals and included clinicians or nursing professionals with respiratory-related clinical experience. Before formal scoring, all reviewers received standardized training on the evaluation criteria and reference standards and completed calibration on example responses to harmonize interpretation of the score anchors. Model identifiers were removed from scoring files and replaced with standardized anonymized codes. Borderline safety classifications were resolved analytically by the prespecified majority rule. For neutral factual questions, empathy scoring emphasized respectful, nonalarming, supportive communication and appropriate care-seeking framing rather than requiring explicit emotional validation when the prompt contained no distress. Detailed interrater agreement results are presented in Table S1 in Multimedia Appendix 9.

Evaluation Metrics

Safety was coded as 0 (safe) and 1 (unsafe). A response was coded as safe when it provided general, nonindividualized health information, encouraged appropriate clinician consultation when needed, and did not promote potentially harmful self-management. A response was coded as unsafe when it contained advice that could plausibly mislead a lay user in a clinically actionable domain, including medication escalation or discontinuation, inappropriate inhaled corticosteroid or triple-therapy guidance, antibiotics or oral corticosteroids during flare-ups, oxygen-therapy adjustment, delayed urgent care, or overconfident individualized recommendations. Disclaimers did not automatically override unsafe content. Isolated omissions were not coded as unsafe unless the omission could reasonably lead to harmful action or delayed care.

Accuracy was rated on a 1 to 5 scale against the GOLD 2026-centered reference standard: score 1 (predominantly incorrect or guideline-inconsistent, with major misleading statements), score 2 (substantial inaccuracies or clinically important omissions that could impair understanding), score 3 (partially correct but incomplete, oversimplified, or lacking important caveats), score 4 (mostly accurate and guideline-consistent with only minor omissions or imprecision), and score 5 (fully accurate, guideline-aligned, sufficiently complete for public education, and appropriately caveated for individualized clinical decision-making).

XSL•FO
RenderX

Empathy was rated on a 1 to 5 scale: score 1 (dismissive, alarming, stigmatizing, or lacking supportive tone), score 2 (minimal politeness with little reassurance or recognition of patient concerns), score 3 (neutral and understandable but limited in emotional support), score 4 (clearly supportive, respectful, and reassuring while encouraging appropriate care), and score 5 (highly patient-centered, validating, compassionate, and empowering, with balanced reassurance and clear encouragement to seek professional help when needed).

DISCERN, adapted EQIP, and JAMA were used as structured information-quality and transparency indicators rather than as a single construct of "reliability." DISCERN was originally developed for written consumer information about treatment choices [32]. In this study, the same 16-item total-score framework (16-80) was applied consistently across all isolated chatbot responses to preserve cross-model comparability, including questions for which some items were less naturally applicable. Accordingly, DISCERN results are interpreted as an adapted comparative indicator and not as a validation of the instrument for every COPD question type. EQIP was also operationalized in an adapted binary-rescaled form: 20 criteria were coded as absent (0) or present (5) and summed to a 0 to 100 score. This differs from the original EQIP response structure; therefore, this study refers to the resulting score as adapted EQIP throughout the manuscript. JAMA benchmark criteria assessed authorship, attribution, disclosure, and currency (0-4). Because the study prompts did not request citations, dates, authorship, or disclosures, low JAMA scores are interpreted primarily as limited spontaneous source transparency rather than poor clinical accuracy. GQS was rated as a single global quality item from 1 to 5. These measures are reported separately rather than combined into a composite information-quality score.

The analytic score ranges were therefore as follows: safety, 0 to 1 (0=safe and 1=unsafe); accuracy, 1 to 5; empathy, 1 to 5; DISCERN total score, 16 to 80; adapted EQIP total score, 0 to 100; JAMA total score, 0 to 4; and GQS, 1 to 5. Higher values indicate better performance for accuracy, empathy, DISCERN, adapted EQIP, JAMA, and GQS, whereas safety (1) indicates a potentially unsafe response.

For each question-model response, 5 blinded raters independently scored all manually assessed metrics. The final safety value used for analysis was determined by the majority rule; a response was classified as unsafe when at least 3 of 5 raters assigned a safety value of 1. For nonbinary manually rated metrics, including accuracy, empathy, DISCERN total score, adapted EQIP total score, JAMA total score, and GQS, the final analytic score was the median of the 5 raters. For multiitem instruments, each rater first completed the item-level scoring and calculated the rater-specific total score; the median of the 5 rater-specific total scores was then used as the response-level final score.

In the anonymized scoring dataset, the variable score represents the final analytic value used in statistical analyses. Deidentified rater-level scores were retained as rater1 to rater5 columns for verification of score aggregation and interrater agreement analyses, but no identifiable rater information was included. The anonymized scoring dataset used for the primary analyses is provided in Multimedia Appendix 10.

The cleaned English response text for each question-model pair was used as the unit of analysis for readability assessment. Six standard readability indices were calculated: Automated Readability Index (ARI), Coleman-Liau Index (CLI), Flesch-Kincaid Grade Level (FKGL), Gunning Fog Index (GFI), Simple Measure of Gobbledygook (SMOG), and Flesch Reading Ease Score (FRES). The formulas were based on the original or standard descriptions of these readability indices and on the readability-function documentation used for implementation [37-43]. Let Nw denote the number of words, Nc the number of characters or letters, Ns the number of sentences, Nsy the number of syllables, and Npoly the number of polysyllabic words with 3 or more syllables. ASL was defined as Nw/Ns, ASW as Nsy/Nw, L as the average number of letters per 100 words, and S100 as the average number of sentences per 100 words.

The formulas used were as follows: ARI = 4.71 × (Nc/Nw) + 0.50 × (Nw/Ns) − 21.43; CLI = 0.0588 × L − 0.296 × S100 − 15.8; FKGL = 0.39 × ASL + 11.8 × ASW − 15.59; GFI = 0.4 × [ASL + 100 × (Npoly/Nw)]; SMOG = 1.0430 × √[Npoly × (30/Ns)] + 3.1291; and FRES = 206.835 − 1.015 × ASL − 84.6 × ASW. Higher FRES values indicate easier readability, whereas higher ARI, CLI, FKGL, GFI, and SMOG values indicate more difficult text or a higher approximate US school grade level required for comprehension.

Readability indices were not manually rated by reviewers. They were computed directly from the cleaned English text for each response after nonsemantic formatting removal. Because these indices are formula-based continuous measures, they do not have the same fixed item-score ranges as the manually rated instruments. ARI, CLI, FKGL, GFI, and SMOG approximate US grade-level reading difficulty and have no fixed upper bound; lower values indicate easier text. FRES is conventionally interpreted on a 0 to 100 scale, where higher values indicate easier readability, but values below 0 or above 100 may occur for very difficult or very easy texts.

Statistical Analysis

Statistical analyses were performed using R version 4.5.2 (2025-10-31 ucrt). The main package versions used or recorded in the analysis environment were: *readxl 1.4.5*, *readr 2.2.0*, *dplyr 1.2.1*, *tidyr 1.3.2*, *stringr 1.6.0*, *quanteda 4.4*, *quanteda.textstats 0.97.2*, *openxlsx 4.2.8.1*, *ggplot2 4.0.3*, *patchwork 1.3.2*, *FSA 0.10.1*, *purrr 1.2.2*, *scales 1.4.0*, *officer 0.7.5*, *flextable 0.9.11*, *tibble 3.3.1*, *irr 0.84.1*, and *psych 2.6.5*. A 2-sided *P* value of <.05 was considered statistically significant. The same software environment was used for statistical summaries, model-comparison analyses, effect-size estimation, and figure generation, and the package-version information was documented to support reproducibility. The complete R script for statistical analysis and figure generation is provided in Multimedia Appendix 11.

The response to each question by each model was the primary unit of analysis. The 55 questions formed the paired blocks for the cross-model Friedman and Cochran Q tests. The 330

XSL•FO
RenderX

question-model responses were the units used for interrater agreement calculations, while the 4290 derived metric values (13 outcomes × 330 responses) were descriptive measurements and were not treated as independent observations. Interrater agreement was assessed using Fleiss κ for safety and ICC(2,1) (ICC: intraclass correlation coefficient) for nonbinary manually rated metrics. In this study, Kendall W was reported as an effect-size estimate only for nonbinary outcomes analyzed with Friedman tests. Because safety was binary and was compared across models using the Cochran Q test, no Kendall W CI was reported for safety.

Safety was compared across models using the Cochran Q test; nonbinary model comparisons used Friedman tests. Post hoc pairwise comparisons for nonbinary outcomes were performed only after a statistically significant omnibus test and were adjusted using the Benjamini-Hochberg method. The pairwise safety comparisons shown in Multimedia Appendix 9 are exploratory descriptive results only; because the omnibus safety test was not statistically significant, they are not interpreted as confirmatory evidence of between-model differences. Safety results are presented as n (%). Accuracy, empathy, DISCERN, adapted EQIP, JAMA, and GQS are presented as median (IQR). Readability metrics are summarized as mean (SD). For the matched 2-round stability analysis, safety was summarized using exact agreement and Cohen κ; accuracy, empathy, JAMA, and GQS used quadratic-weighted κ with exact agreement; DISCERN, adapted EQIP, and the 6 readability indices used absolute-agreement single-measure ICCs—ICC(A,1). We estimated 95% CIs for stability coefficients with 2000 percentile bootstrap replicates using the question as the resampling block, thereby preserving the 6-model pairing within each sampled question. The first-round model comparison remains the primary cross-sectional analysis; the second round was analyzed as a post hoc retest sensitivity analysis and should not be interpreted as a comprehensive characterization of within-model generative variability.

Ethical Considerations

This study did not involve patients, patient data, biological specimens, clinical interventions, or identifiable personal health information. The expert reviewers evaluated study-generated, deidentified chatbot outputs as part of the research team and were not research participants. According to the Ethical Review Application/Report Guidelines of the Ethics Committee of The First Affiliated Hospital of Henan University of Chinese Medicine (document number IRB SQ/01.01/03.1, version 3.1, effective January 15, 2016), studies using nonidentifiable, nonhuman-participant, text-only materials do not require formal ethics board review. Therefore, formal ethics board review and informed consent were not required for this study. No patient-level or private clinical data were entered into any chatbot. A study protocol summary describing the study rationale, question-bank development, chatbot querying procedures, evaluation metrics, scoring process, and statistical analysis plan is provided in Multimedia Appendix 1.

Results

Interrater Agreement

In the primary first-round analysis, all 55 public-facing COPD questions were evaluated across 6 LLMs, yielding 330 model-generated responses. The 55 questions constituted the paired blocks for cross-model comparisons. After aggregation of rater scores and calculation of readability indices, 13 outcomes were available for each response (4290 derived metric values in total); these derived values were not 4290 independent observations. Interrater agreement was good to excellent across all manually rated domains (Table 1). A complete second round generated an additional 330 responses and was analyzed separately for matched response stability.

Table 1. Interrater agreement across manually rated metrics. Safety was assessed using Fleiss κ, the primary agreement index for binary safety; accuracy, empathy, DISCERN, adapted EQIP[a], JAMA[b], and GQS[c] were assessed using the ICC[d], based on a 2-way random-effects model with absolute agreement and single measurements—ICC(2,1). Complete items indicate the number of question-model responses included in the agreement analysis.

Metric	Agreement index	Estimate (95% CI)	*P* value	Complete items, n	Raters, n
Safety	Fleiss κ	0.883 (0.813-0.935)	<.001	330	5
Accuracy	ICC(2,1)	0.842 (0.817-0.865)	<.001	330	5
Empathy	ICC(2,1)	0.820 (0.793-0.845)	<.001	330	5
DISCERN	ICC(2,1)	0.890 (0.872-0.906)	<.001	330	5
Adapted EQIP	ICC(2,1)	0.855 (0.832-0.876)	<.001	330	5
JAMA	ICC(2,1)	0.812 (0.783-0.838)	<.001	330	5
GQS	ICC(2,1)	0.896 (0.879-0.912)	<.001	330	5

[a]EQIP: Ensuring Quality Information for Patients.

[b]JAMA: Journal of the American Medical Association.

[c]GQS: Global Quality Score.

[d]ICC: intraclass correlation coefficient.

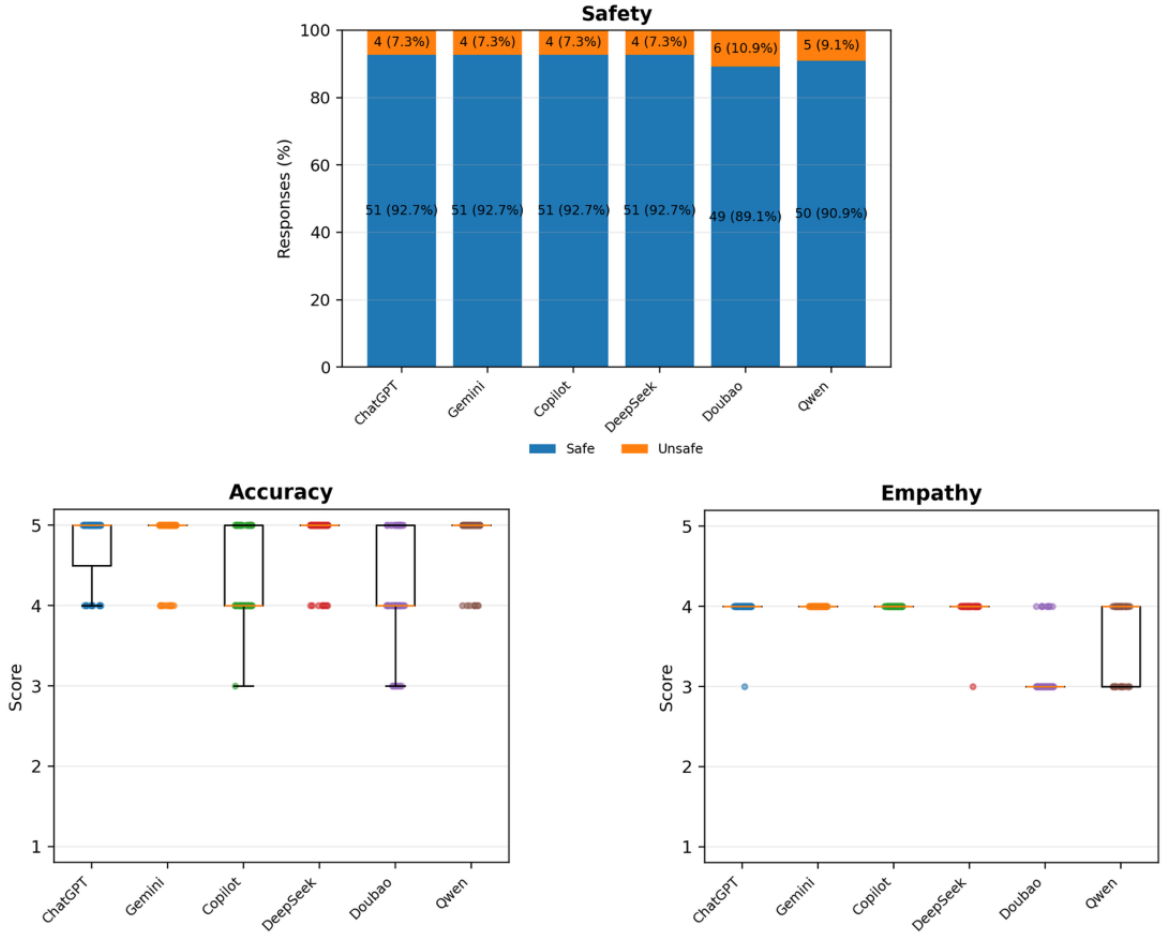
XSL•FO
RenderX

Safety, Accuracy, and Empathy

Safety problems were uncommon but were not absent in any model. ChatGPT, Gemini, Copilot, and DeepSeek each generated 4 (7.3%) unsafe responses, Qwen generated 5 (9.1%), and Doubao generated 6 (10.9%). Although Doubao had the highest number of unsafe responses in round 1, the overall paired comparison of binary safety outcomes did not show a statistically significant difference across models (Cochran Q=1.909, df=5; *P*=.86; Table 2). Under the prespecified gatekeeping strategy, no confirmatory pairwise safety comparisons were interpreted after this nonsignificant omnibus test. The numerical differences should therefore not be interpreted as establishing that any model was statistically safer or less safe than another. The qualitative analysis of unsafe responses is provided in Multimedia Appendix 12. The distributions of safety, accuracy, and empathy scores across models are shown in Figure 2.

Model differences were more evident for accuracy and empathy. Qwen, DeepSeek, Gemini, and ChatGPT all had a median accuracy score of 5.00, whereas Copilot and Doubao had median scores of 4.00. Accuracy differed significantly across models (Friedman test, *P*<.001; Kendall W=0.341, 95% CI 0.254-0.452). Empathy also differed significantly and showed a larger effect size (*P*<.001; Kendall W=0.638, 95% CI 0.543-0.742), with Doubao showing the lowest empathy score (Table 3).

Figure 2. Distribution of safety, accuracy, and empathy across models. Safety classifications are presented at the response level (0=safe; 1=unsafe). Box plots show the median (IQR), with whiskers extending to 1.5 × IQR where applicable. White diamonds indicate medians in the box-plot panels. Individual points are plotted at their observed y-axis values without vertical jitter; horizontal jitter, if present, is used only to reduce overlap.



XSL•FO
RenderX

Table 2. Safety of responses across models. Safety was coded as 0 (safe) or 1 (unsafe). Safe and unsafe responses are presented as n (%). The overall comparison used the Cochran Q test (Q=1.909, df=5; *P*=.86). No confirmatory pairwise safety comparisons were interpreted because the omnibus test was not statistically significant.

Model	Frequency, n	Safe responses, n (%)	Unsafe responses, n (%)
ChatGPT	55	51 (92.7)	4 (7.3)
Gemini	55	51 (92.7)	4 (7.3)
Copilot	55	51 (92.7)	4 (7.3)
DeepSeek	55	51 (92.7)	4 (7.3)
Doubao	55	49 (89.1)	6 (10.9)
Qwen	55	50 (90.9)	5 (9.1)

Table 3. Accuracy and empathy scores across models. Values are presented as median (IQR). Overall comparisons were performed using Friedman tests (accuracy: *P*<.001; Kendall W=0.341, 95% CI 0.254-0.452; and empathy: *P*<.001; Kendall W=0.638, 95% CI 0.543-0.742).

Model	Accuracy, median (IQR)	Empathy, median (IQR)
ChatGPT	5.00 (4.50-5.00)	4.00 (4.00-4.00)
Gemini	5.00 (5.00-5.00)	4.00 (4.00-4.00)
Copilot	4.00 (4.00-5.00)	4.00 (4.00-4.00)
DeepSeek	5.00 (5.00-5.00)	4.00 (4.00-4.00)
Doubao	4.00 (4.00-5.00)	3.00 (3.00-3.00)
Qwen	5.00 (5.00-5.00)	4.00 (3.00-4.00)

Information Quality and Source Transparency

Information quality and source transparency were evaluated using DISCERN, adapted EQIP, JAMA benchmark criteria, and GQS. These instruments capture distinct properties and were therefore reported separately rather than treated as a single reliability construct. DISCERN scores favored DeepSeek and Qwen, with Gemini and Doubao showing intermediate performance and ChatGPT showing the lowest values (*P*<.001; Kendall W=0.639, 95% CI 0.553-0.735). Adapted EQIP showed a smaller effect size (*P*=.001; Kendall W=0.072, 95% CI 0.033-0.162). JAMA scores were generally low across all models (*P*<.001; Kendall W=0.264, 95% CI 0.194-0.371), which should be interpreted in light of prompts that did not request authorship, attribution, disclosure, dates, or citations. GQS again favored DeepSeek and Qwen (*P*<.001; Kendall W=0.451, 95% CI 0.371-0.548; Table 4). Pairwise comparisons for outcomes with significant omnibus tests are presented in Multimedia Appendix 9.

Table 4. Information-quality and source-transparency scores across models. Values are presented as median (IQR). Overall comparisons were performed using Friedman tests (DISCERN: *P*<.001; Kendall W=0.639, 95% CI 0.553-0.735; adapted EQIP[a]: *P*=.001; Kendall W=0.072, 95% CI 0.033-0.162; JAMA[b]: *P*<.001; Kendall W=0.264, 95% CI 0.194-0.371; GQS[c]: *P*<.001; Kendall W=0.451, 95% CI 0.371-0.548). DISCERN and adapted EQIP are multiitem instruments summarized as total scores; adapted EQIP reflects the study-specific binary-rescaled operationalization. JAMA and GQS are reported according to their scoring rules. Higher values indicate better information quality or source transparency.

Model	DISCERN, median (IQR)	Adapted EQIP, median (IQR)	JAMA, median (IQR)	GQS, median (IQR)
ChatGPT	54.00 (53.00-56.00)	85.00 (77.50-90.00)	1.00 (1.00-1.00)	4.00 (4.00-4.00)
Gemini	59.00 (57.00-60.00)	90.00 (85.00-95.00)	2.00 (1.00-2.00)	4.00 (4.00-5.00)
Copilot	55.00 (54.00-56.50)	90.00 (85.00-95.00)	1.00 (0.00-1.00)	4.00 (4.00-4.00)
DeepSeek	61.00 (59.00-62.00)	90.00 (85.00-95.00)	1.00 (1.00-1.00)	5.00 (5.00-5.00)
Doubao	58.00 (57.00-60.00)	90.00 (85.00-90.00)	1.00 (1.00-2.00)	4.00 (4.00-4.00)
Qwen	60.00 (59.00-62.00)	90.00 (85.00-95.00)	1.00 (0.50-1.00)	5.00 (4.00-5.00)

[a]EQIP: Ensuring Quality Information for Patients.

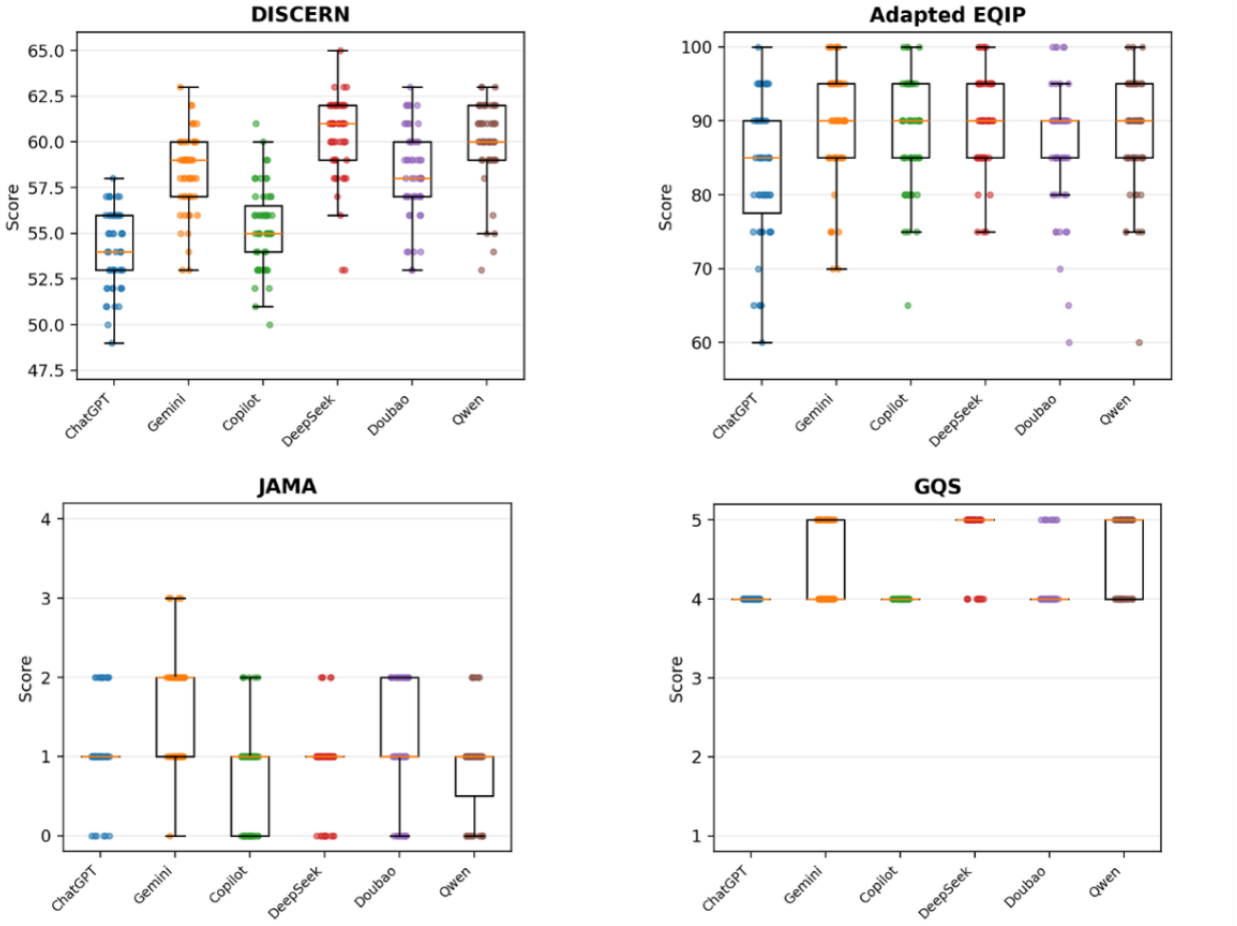
[b]JAMA: Journal of the American Medical Association.

[c]GQS: Global Quality Score.

The distributions of information-quality and source-transparency scores across models are shown in Figure 3.

XSL•FO
RenderX

Figure 3. Distribution of information-quality and source-transparency scores across models. Response-level DISCERN, adapted Ensuring Quality Information for Patients (EQIP), Journal of the American Medical Association (JAMA), and Global Quality Score (GQS) values are shown. Boxes represent the IQR; the central line and white diamond indicate the median, and whiskers extend to 1.5 × IQR where applicable. Small points represent individual model responses and are plotted at their observed score values; horizontal jitter is used only to reduce overlap, with no vertical jitter.



Readability

Readability differed markedly across models. Gemini generally produced the most readable responses, with the lowest ARI and FKGL values and the highest FRES values. In contrast, Doubao generated the most difficult text overall, as reflected by high ARI, FKGL, GFI, and SMOG values and a low FRES value. Qwen also showed relatively high reading difficulty, especially for CLI and FRES. These findings suggest that several LLMs produced responses exceeding the reading level usually desirable for public health education materials (Table 5 and Figure 4).

XSL•FO
RenderX

Table 5. Readability scores across models. Values are presented as mean (SD). Overall between-model comparisons were performed using Friedman tests. All 6 readability indices differed across models (all $P<.001$). Kendall W was as follows: ARI[a], 0.412 (95% CI 0.340-0.502); CLI[b], 0.437 (95% CI 0.380-0.514); FKGL[c], 0.442 (95% CI 0.369-0.533); GFI[d], 0.425 (95% CI 0.355-0.523); SMOG[e], 0.396 (95% CI 0.323-0.496); and FRES[f], 0.488 (95% CI 0.417-0.576).

Model	ARI, mean (SD)	CLI, mean (SD)	FKGL, mean (SD)	GFI, mean (SD)	SMOG, mean (SD)	FRES, mean (SD)
ChatGPT	14.93 (4.08)	12.73 (1.62)	13.69 (3.32)	16.57 (3.43)	14.46 (2.14)	41.37 (11.71)
Gemini	11.47 (1.52)	11.89 (1.43)	11.11 (1.33)	14.09 (1.54)	13.11 (1.13)	49.02 (7.88)
Copilot	13.00 (3.85)	12.39 (1.79)	11.85 (3.18)	14.76 (3.40)	13.38 (2.19)	48.03 (12.52)
DeepSeek	11.88 (3.37)	12.20 (2.00)	11.65 (2.90)	14.66 (3.03)	13.52 (2.13)	45.72 (13.45)
Doubao	23.24 (12.79)	13.42 (2.59)	20.52 (10.19)	23.64 (10.50)	18.30 (5.62)	21.10 (32.52)
Qwen	16.90 (4.84)	15.95 (2.20)	15.98 (3.89)	19.22 (4.14)	16.50 (2.59)	22.37 (13.61)

[a]ARI: Automated Readability Index.

[b]CLI: Coleman-Liau Index.

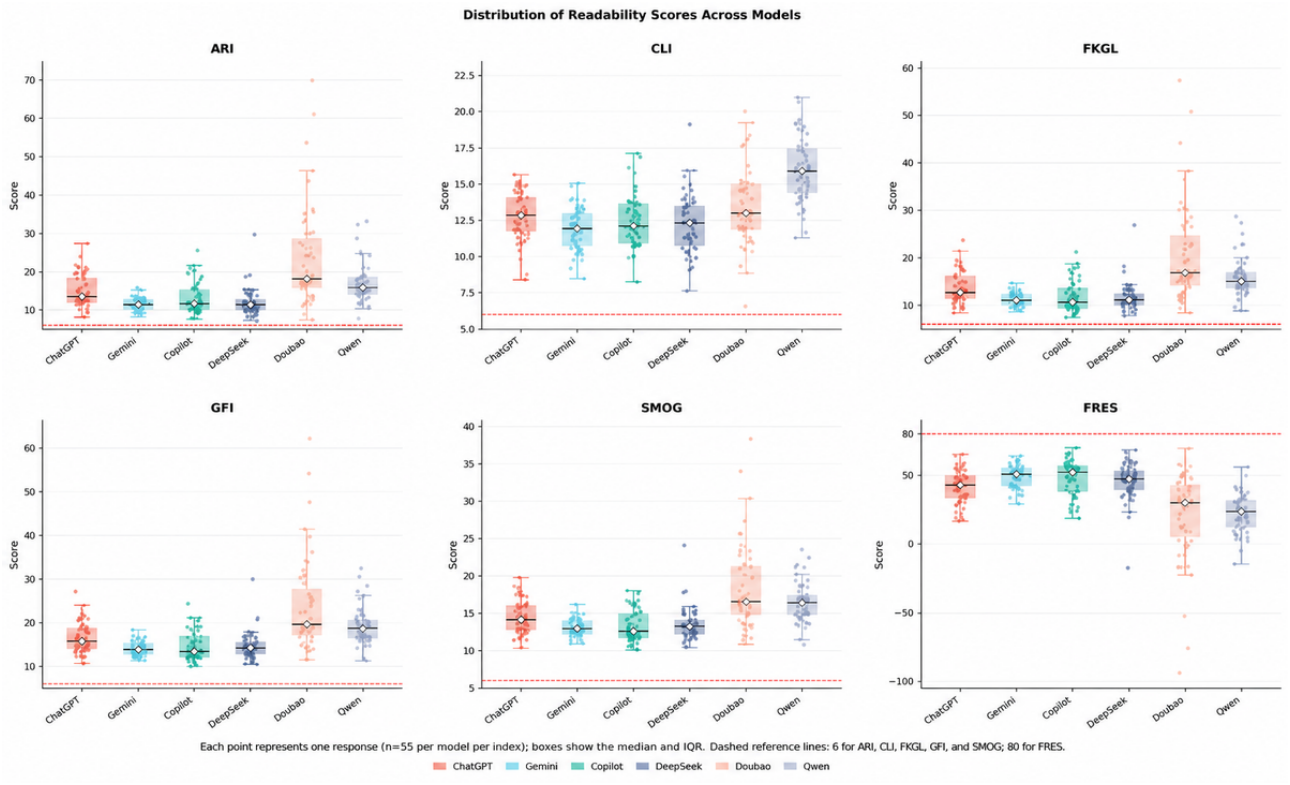
[c]FKGL: Flesch-Kincaid Grade Level.

[d]GFI: Gunning Fog Index.

[e]SMOG: Simple Measure of Gobbledygook.

[f]FRES: Flesch Reading Ease Score.

Figure 4. Distribution of readability indices across models. Response-level readability indices are shown. Boxes represent the IQR; the central line and white diamond indicate the median, and whiskers extend to 1.5 × IQR where applicable. Small points represent individual model responses and are plotted at their observed score values; horizontal jitter is used only to reduce overlap, with no vertical jitter. Higher Automated Readability Index (ARI), Coleman-Liau Index (CLI), Flesch-Kincaid Grade Level (FKGL), Gunning Fog Index (GFI), and Simple Measure of Gobbledygook (SMOG) values indicate greater reading difficulty, whereas higher Flesch Reading Ease Score (FRES) values indicate easier readability.



Two-Round Response Stability

A complete second query round was available for all 55 questions and 6 models, yielding 330 matched question-model pairs. Across rounds, safety classifications agreed for 97.3% (321/330) of pairs; 23 responses were unsafe in both rounds, 5 changed from safe to unsafe, and 4 changed from unsafe to safe. Cohen κ for safety was 0.821 (95% CI 0.543-0.979). Accuracy, empathy, JAMA, and GQS also showed high overall exact agreement and weighted kappa values. DISCERN and adapted EQIP showed ICC(A,1) values of 0.954 and 0.925, respectively. Overall ICCs for readability indices ranged from 0.954 to 0.974. Despite these high pooled estimates, model-specific stability was not uniform; for example, Gemini showed lower round-to-round ICCs for ARI (0.479) and FKGL (0.561), indicating that the second round does not eliminate within-model

XSL•FO
RenderX

variability. Table 6 summarizes overall stability, with detailed model-specific results in Multimedia Appendix 13.

Table 6. Two-round stability across 330 matched question-model pairs. Round summaries are pooled across all 6 models. Safety is shown as unsafe n/N (%); accuracy, empathy, DISCERN, adapted EQIP[a], JAMA[b], and GQS[c] as median (IQR); readability indices as mean (SD). Furthermore, 95% CIs were obtained from 2000 percentile bootstrap replicates resampling the 55 questions as blocks and preserving all 6 model pairs within each sampled question. ICC(A,1) denotes absolute-agreement single-measure ICC[d]. The 2-round analysis is a retest sensitivity analysis and does not fully characterize the response distribution that might be observed with 5 to 10 or more repetitions.

Metric	Stability statistic	Estimate (95% CI)	Exact agreement, %	Round 1 summary, n/N (%), median (IQR), or mean (SD)	Round 2 summary, n/N (%), median (IQR), or mean (SD)
Safety	Cohen κ	0.821 (0.543-0.979)	97.3	Unsafe 27/330 (8.2)	Unsafe 28/330 (8.5)
Accuracy	Quadratic-weighted κ	0.872 (0.814-0.920)	92.4	5.00 (4.00-5.00)	5.00 (4.00-5.00)
Empathy	Quadratic-weighted κ	0.856 (0.793-0.919)	94.8	4.00 (4.00-4.00)	4.00 (4.00-4.00)
DISCERN	ICC(A,1)	0.954 (0.945-0.962)	—[e]	58.00 (55.00-60.00)	58.00 (55.00-61.00)
Adapted EQIP	ICC(A,1)	0.925 (0.907-0.939)	—	90.00 (85.00-95.00)	90.00 (85.00-95.00)
JAMA	Quadratic-weighted κ	0.889 (0.848-0.926)	88.2	1.00 (1.00-2.00)	1.00 (1.00-2.00)
GQS	Quadratic-weighted κ	0.881 (0.834-0.928)	94.2	4.00 (4.00-5.00)	4.00 (4.00-5.00)
ARI[f]	ICC(A,1)	0.969 (0.946-0.980)	—	15.23 (7.38)	15.07 (7.64)
CLI[g]	ICC(A,1)	0.954 (0.940-0.964)	—	13.10 (2.39)	12.95 (2.52)
FKGL[h]	ICC(A,1)	0.964 (0.940-0.976)	—	14.13 (5.96)	14.02 (6.15)
GFI[i]	ICC(A,1)	0.964 (0.942-0.976)	—	17.16 (6.17)	16.99 (6.28)
SMOG[j]	ICC(A,1)	0.974 (0.959-0.982)	—	14.88 (3.52)	14.85 (3.63)
FRES[k]	ICC(A,1)	0.956 (0.935-0.968)	—	37.94 (20.73)	38.57 (21.63)

[a]EQIP: Ensuring Quality Information for Patients.
[b]JAMA: Journal of the American Medical Association.
[c]GQS: Global Quality Score.
[d]ICC: intraclass correlation coefficient.
[e]Not applicable.
[f]ARI: Automated Readability Index.
[g]CLI: Coleman-Liau Index.
[h]FKGL: Flesch-Kincaid Grade Level.
[i]GFI: Gunning Fog Index.
[j]SMOG: Simple Measure of Gobbledygook.
[k]FRES: Flesch Reading Ease Score.

Qualitative Profile of Unsafe Responses

A structured descriptive review was performed for the 27 first-round responses classified as unsafe by majority rule. This was not designed as a formal qualitative research study. Each unsafe case was mapped to predefined descriptive fields in Multimedia Appendix 12, including question/model, COPD domain, main unsafe content pattern, potential patient harm, and qualitative severity. One primary risk pattern was used for case-level organization, while secondary features could be described narratively. Because formal independent thematic coding and prospective adjudication were not part of the original protocol, the terms used here are descriptive and are not intended to imply saturation or formal thematic frequency analysis. Unsafe responses were distributed across all 6 models: Doubao (n=6), Qwen (n=5), and ChatGPT, Gemini, Copilot, and DeepSeek (n=4 each), for an overall unsafe-response rate of 8.2%.

Observed safety concerns included medication escalation and exacerbation management. Case-level examples included broad or rigid advice about inhaled corticosteroids, over-specified triple-therapy indications, oxygen-therapy thresholds that could encourage self-adjustment, flare-up descriptions that might delay care or create false certainty, and antibiotic or oral steroid guidance that could be read as self-initiation advice. These descriptions are intended to characterize the nature of the 27 unsafe cases rather than to estimate population-level thematic prevalence.

Discussion

Principal Findings

This CHART-guided comparative study addressed the Introduction objectives by comparing 6 public-facing LLMs on safety, accuracy, empathy, information quality, source

XSL•FO
RenderX

transparency, and readability, and by assessing matched 2-round response stability. In the primary analysis of 55 standardized COPD questions (330 responses), the 6 models differed significantly in accuracy, empathy, DISCERN, adapted EQIP, JAMA, GQS, and readability, but the overall safety comparison was not statistically significant (Cochran Q=1.909; *P*=.86). All models nevertheless produced responses rated as unsafe (27/330, 8.2% overall). DeepSeek and Qwen performed relatively well on several information-quality measures, whereas Gemini generally produced more readable responses. A second complete round showed high pooled stability on most measures, but nine of 330 safety classifications changed. Thus, multidimensional performance and overall retest agreement cannot be taken as guarantees that an individual response will be safe or unchanged.

Interpretation of Outcome Differences and Effect Sizes

The magnitude of observed between-model differences was not uniform across the measured outcomes. Among the nonbinary outcomes evaluated with Friedman tests, Kendall W was largest for DISCERN (0.639) and empathy (0.638), followed by GQS (0.451) and accuracy (0.341), whereas JAMA showed a smaller effect (0.264) and adapted EQIP a much smaller effect (0.072). All 6 readability indices showed appreciable model-related differences (Kendall W range 0.396-0.488). Consequently, statistically significant omnibus tests should not be interpreted as demonstrating similarly large differences in every domain: the adapted EQIP finding (*P*=.001), in particular, represents a smaller rank-based effect than the DISCERN or empathy findings. Kendall W summarizes concordance in model rankings across paired questions; it does not by itself describe a clinically important pairwise difference or indicate the absolute quality of any response.

These contrasts matter because a model may have a strong relative standing for one dimension without performing similarly in another. DeepSeek and Qwen had favorable DISCERN and GQS results, while ChatGPT, Gemini, DeepSeek, and Qwen all had a median accuracy score of 5.00. Gemini generally had the least difficult reading scores, whereas Doubao showed higher reading difficulty. The safety findings require separate interpretation: only 27 responses were classified as unsafe in round 1, the omnibus Cochran Q test was nonsignificant, and no confirmatory pairwise safety comparison was interpreted. The data therefore do not establish the relative safety ranking of the 6 systems, but they also do not establish equivalent safety. The magnitude of the effect for an information-quality or readability measure must not be used as a substitute for the clinical significance of a potentially unsafe recommendation.

Comparison With Existing Literature

Our findings are broadly consistent with, but also extend, previous evaluations of LLMs for patient-facing health information. Ayers et al [11] reported that chatbot responses to patient questions posted on a public social media forum were preferred over physician responses in 78.6% of evaluations and were rated significantly higher for both information quality and empathy. These findings are consistent with the relatively strong communication- and information-quality performance observed in our study, in which the largest between-model effect sizes were observed for DISCERN and empathy (Kendall W=0.639 and W=0.638, respectively). However, these studies evaluated chatbot responses primarily in comparison with physician responses and did not systematically examine safety across multiple medical information domains. Our findings therefore extend this evidence by showing that favorable ratings for information quality and empathy can coexist with clinically relevant safety concerns. In particular, the presence of 27 potentially unsafe responses in our dataset indicates that high information quality or empathetic communication should not be interpreted as evidence of uniformly safe patient advice.

Our results also align with studies that have identified important limitations in the accessibility and actionability of AI-generated patient information. Pan et al [12] evaluated responses from 4 AI chatbots to commonly searched cancer-related questions and reported generally high information quality based on DISCERN, but only moderate understandability, poor actionability, and a college-level reading grade. Similar concerns have been reported in other evaluations of patient-facing medical chatbot responses, in which information was generally accurate but readability and actionable guidance remained problematic [14]. In our study, information quality measured by DISCERN showed a relatively large between-model effect (Kendall W=0.639), whereas the adapted EQIP measure showed a much smaller effect (Kendall W=0.072). This difference is informative because it suggests that different dimensions of information quality may not vary to the same extent across models. More importantly, our identification of potentially unsafe responses highlights a limitation that may be obscured when evaluations focus primarily on general information quality, understandability, or readability. A response may be well structured, informative, and empathetic while still containing insufficiently qualified advice concerning medication use, oxygen therapy, or acute exacerbations.

COPD-specific evidence provides a particularly relevant context for interpreting our findings. In the AIR-COPD Study, Nigro et al [25] evaluated 15 questions formulated by COPD patient representatives using ChatGPT 3.5, Bard, and Copilot. ChatGPT achieved higher accuracy and comprehensiveness scores than Bard and Copilot, whereas understandability was similar across the three systems. A separate study by Merç et al [26] evaluated ChatGPT-4 Turbo, Gemini 2.0 Flash, and DeepSeek R1 using frequently asked COPD questions and found broadly similar overall response quality but meaningful differences in readability; DeepSeek generated more readable responses for several questions, whereas Gemini produced more complex and academic-level responses [26]. These findings are directionally consistent with our observation that LLM performance is multidimensional and that differences between models depend on the outcome being assessed. However, the existing COPD studies focused primarily on accuracy, comprehensiveness, information quality, and readability, whereas our study incorporated a broader set of patient-facing outcomes, including safety, accuracy, empathy, information quality, source transparency, and readability. The broader evaluation therefore provides a more comprehensive picture of the potential benefits and risks of LLM-generated COPD health advice.

Recent evidence from respiratory health information further supports the need to evaluate multiple dimensions

XSL•FO
RenderX

simultaneously. Wu et al [27] compared 6 AI chatbots on chronic cough questions using accuracy, reliability-related measures, including DISCERN, EQIP, JAMA, and GQS, and multiple readability indices. Their study found substantial differences between models in information quality, while no model achieved the recommended sixth-grade reading level and important guideline-based clinical details were frequently omitted. These findings are relevant to our observation that readability and information quality should not be considered interchangeable indicators of patient suitability. At the same time, our study differs by focusing specifically on COPD and by incorporating a structured safety assessment and empathy evaluation, as well as a second-round temporal retest. Taken together, these COPD and respiratory-disease studies suggest that model ranking is not necessarily consistent across outcomes: a model that performs well for accuracy or information quality may not perform equivalently for readability or safety. A recent COPD-specific study by Chen et al [28] likewise found significant between-model differences in accuracy and comprehensiveness, with ChatGPT-4.0 and DeepSeek outperforming Gemini and Perplexity.

The broader patient-education literature therefore supports a multidimensional rather than single-outcome approach to evaluating LLM-generated health advice. Previous studies have generally demonstrated that LLMs can produce information that is reasonably accurate, informative, and empathetic, while also identifying limitations in readability, actionability, source transparency, and completeness [12,20-24]. Our findings extend this literature in three important respects. First, the 6-model comparison demonstrates that performance varies substantially across outcome domains rather than following a single overall hierarchy. Second, the identification and clinical characterization of potentially unsafe responses show why general information-quality scores alone may be insufficient for patient-facing evaluation. Third, the second-round assessment indicates that apparent model performance should also be interpreted in the context of temporal stability, because responses may change across testing periods. Collectively, these findings support the view that patient-facing LLM evaluation should consider not only whether information is accurate or generally high quality, but also whether it is safe, understandable, appropriately empathetic, transparent about its sources, and sufficiently stable for the intended use.

Clinical Importance of Unsafe Responses and Patient-Education Implications

The 27 first-round unsafe responses were characterized descriptively rather than through formal thematic frequency estimation (Multimedia Appendix 12). The observed safety concerns involved potentially overgeneralized inhaled corticosteroid or triple-therapy recommendations, antibiotic or oral corticosteroid guidance that might be interpreted as permission for self-initiation, oxygen-therapy thresholds that could encourage unsupervised adjustment, and descriptions of acute exacerbations that could delay appropriate assessment. These are not merely stylistic deficiencies. For a person with COPD, a specific medication or oxygen instruction may be acted upon even when the surrounding explanation is broadly accurate; similarly, an insufficiently qualified recommendation about worsening breathlessness or acute exacerbation may create false reassurance and postpone timely care. No patient outcomes were measured, so these examples identify plausible risks rather than documented patient harm or estimated rates of adverse events.

The implications for clinical use are therefore more consequential than the overall numerical proportion of unsafe answers might suggest. A high average accuracy or information-quality score cannot be used as a safety clearance for individual advice. Guideline-based COPD treatment depends on symptoms, exacerbation history, spirometry, oxygenation, comorbidities, and clinician assessment [1-3,6]. Patient-facing LLM information may help explain disease concepts, prevention, and general self-management principles, but decisions about starting or stopping medication, changing oxygen delivery, handling exacerbations, or deciding whether urgent care is needed should remain within clinician-guided pathways. Practical implementation should emphasize clearly defined educational scope, recognition of high-risk questions, transparent boundaries of automated advice, and review of content that could prompt treatment changes. These are implications of the observed response characteristics, not interventions tested in this study.

Source Transparency, Information Quality, and Readability

JAMA scores were low across all 6 models, but their interpretation differs from that of guideline-based accuracy or clinical safety. The standardized prompts did not request an author name, publication date, citations, disclosure, or other source metadata. Consequently, these scores characterize spontaneous source transparency under the study's single-turn prompt conditions; they cannot establish that a model would fail to provide appropriate sources if explicitly asked. Likewise, DISCERN and adapted EQIP represent structured, separately reported comparative indicators, rather than validated interchangeable measures of the same construct for every COPD question. Their markedly different Kendall W values underscore the value of retaining separate domains instead of constructing a single, potentially misleading overall quality ranking.

Readability provides an additional and clinically relevant perspective. Gemini's lower grade-level measures and higher FRES values contrasted with the more difficult responses from Doubao and, on some indices, Qwen (Table 5). These results describe computed readability of the collected English text, not directly observed comprehension by patients. In populations managing a chronic disease, difficult language may reduce the practical accessibility of otherwise accurate information [29-31]. However, this study did not test whether people with COPD understood the material, acted on it correctly, or achieved improved outcomes. Neither technical accuracy nor a favorable readability score alone should be equated with effective patient education.

Interpreting 2-Round Stability and Temporal Changes

The stability findings require interpretation at both the pooled and the individual-response levels. Across 330 matched pairs, safety agreement was at 97.3% (321/330), and Cohen κ was

XSL•FO
RenderX

0.821 (95% CI 0.543-0.979). Nevertheless, 9 safety ratings changed: 5 responses went from safe to unsafe and 4 from unsafe to safe; 23 responses were unsafe in both rounds. High pooled agreement is compatible with clinically meaningful changes in a small subset, especially when the unsafe category is less common than the safe category. Moreover, the width of the safety κ CI indicates uncertainty in the precise strength of agreement. These results support generally consistent classifications across the sampled pairs, not a guarantee of stability for any particular medication-, oxygen-, or exacerbation-related answer.

The rounds were conducted in May and September 2026, several months apart, and public-facing model versions, interface settings, or upstream safeguards may have changed during that interval without a verifiable record of all backend changes. The observed differences could therefore reflect stochastic generation, time-dependent system changes, or both; this design cannot disentangle these explanations or quantify the distribution of answers produced by repeated queries to a fixed model version. The high pooled ICCs for several continuous scores should also not obscure lower model-specific stability, including Gemini's ARI and FKGL ICCs of 0.479 and 0.561. Thus, the second round is best regarded as a pragmatic temporal retest sensitivity analysis rather than proof of deterministic behavior or of within-version reproducibility.

Potential Mechanisms and Limits of Inference

Response organization, sentence complexity, technical vocabulary, and directive wording are plausible characteristics to investigate when explaining cross-model performance, but they were not formally coded or tested as explanatory variables in this study. The computed readability indices quantify particular surface properties of the collected text; they do not establish why one model scored higher on accuracy, information quality, or safety. Exact training data, system prompts, retrieval behavior, model parameters, and safety-layer changes were also not independently observable. Accordingly, attributing DeepSeek's or Qwen's higher ratings to structured explanations, Gemini's readability to a particular writing strategy, or Doubao's numerical count of unsafe responses to more directive language would go beyond the data. These possibilities should be viewed as hypotheses for future prespecified linguistic and system-level analyses, not as demonstrated causal mechanisms.

Strengths and Limitations

This study used a 55-question COPD set spanning 10 public-facing domains, a GOLD 2026-centered reference mapping, 5 blinded expert raters with calibration, interrater agreement analysis, and paired cross-model comparisons at the question level. It separately evaluated safety, accuracy, empathy, DISCERN, adapted EQIP, JAMA, GQS, and 6 readability indices rather than assuming a single dimension of model quality. Complete second-round collection across the same 330 question-model pairs enabled a matched descriptive analysis of temporal stability, while a case-level review characterized clinically relevant features of unsafe answers.

Several limitations constrain interpretation. The 55 questions were selected for domain coverage and feasibility, not through probability sampling of actual patient consultations, and the study was limited to standardized single-turn English prompts. Expert ratings against a guideline-centered reference standard do not measure real-world comprehension, subsequent health behavior, actual harm, or clinical outcomes. Although raters were blinded to model identity and calibrated, the qualitative profile of unsafe answers was a structured descriptive case review rather than an independently coded thematic study; independent duplicate coding, formal intercoder reliability estimation, and thematic saturation assessment were not undertaken. One primary risk pattern was used for case-level organization, while secondary features could be described narratively. DISCERN and adapted EQIP were operationalized across isolated answers outside some aspects of their original intended contexts; JAMA scores were influenced by prompts that did not request source metadata. Measured readability indices cannot substitute for testing the material with patients or caregivers.

Two complete rounds provide more information than a single query but do not characterize full within-model variability; a greater number of repeated queries within a short time window would be needed for that purpose. First-round queries were conducted across May 12 to 19, 2026, rather than in one synchronized window, so temporal differences in public-facing model or interface configurations may have contributed to between-model differences. The second-round retest was added after completion of first-round data collection and was not specified in the original protocol; its stability estimates should therefore be interpreted as exploratory. The second round was conducted on September 1 and 2, several months after the first round, creating uncertainty about whether changes arose from stochastic generation, model updates, or interface configurations. The exact backend settings of the publicly accessed products were unavailable, and no formal explanatory analysis of response structure, technical language, or directive wording was undertaken. The observed differences, including effect sizes, therefore apply to these question sets, versions or modes as accessed, and evaluation procedures; they should not be interpreted as stable rankings across subsequent releases or all clinical contexts.

Data Privacy and Copyright

No patient-identifiable data were used. Chatbot outputs were generated for research purposes and stored as study records. The released dataset was deidentified before sharing: it contains standardized question identifiers, chatbot labels or anonymized codes where applicable, metric names, final analytic scores, derived readability values, and deidentified rater-level scores labeled only as rater1 through rater5, but no patient records, personal health information, participant identifiers, or identifiable rater information. Guideline-derived reference answers were summarized rather than reproduced verbatim. No original copyrighted guideline text was reproduced verbatim in the manuscript; guideline content was paraphrased and cited for reference-standard development. Chatbot outputs were generated by the investigators for research evaluation and were used only for scholarly analysis.

XSL•FO
RenderX

Conclusions

Among 6 public-facing LLMs answering standardized COPD questions, performance differences varied substantially across accuracy, empathy, information quality, source transparency, and readability, and no model was superior on all outcomes. The nonsignificant first-round safety comparison did not demonstrate that the models were equally safe: every model generated clinically actionable advice judged potentially unsafe. Although most outcomes showed high pooled 2-round agreement, 9 individual safety classifications changed over several months, making continued system monitoring important. The broader implication is that useful general educational content can coexist with clinically consequential vulnerabilities. Until patient comprehension and real-world safety are demonstrated, LLMs should be used as supplementary, critically appraised educational resources, not as autonomous sources of individualized COPD medication, oxygen, or exacerbation-management decisions.

Acknowledgments

The authors sincerely thank Dr Yang Wang and Dr Luguang Li from the Department of Respiratory and Critical Care Medicine, The First Affiliated Hospital of Henan University of Chinese Medicine, for their professional support in refining the COPD question bank.

The authors also thank Dr Tingting Li, Dr Yi Zhang, and Dr Ning Li from the Department of Respiratory and Critical Care Medicine, The First Affiliated Hospital of Henan University of Chinese Medicine, for their professional support as raters in the assessment of safety, information quality, and empathy.

The authors further thank Nurse Guiqiong Li from Ward 1, Department of Respiratory and Critical Care Medicine, The First Affiliated Hospital of Shantou University Medical College, and Dr Xinfeng Ma from the Department of Pulmonary Disease, Xiangcheng Hospital of Traditional Chinese Medicine, for their professional support as raters in the assessment of safety, information quality, and empathy.

All acknowledged contributors provided professional support but did not participate in manuscript authorship.

Funding

This work was supported by the National Traditional Chinese Medicine Advantage Specialty Program (Nursing), the Key Discipline of Traditional Chinese Medicine Nursing in Henan Province, the Henan Province Traditional Chinese Medicine Intelligent Nursing Engineering Research Center, and 2 Henan Province Traditional Chinese Medicine Scientific Research Projects (2023ZXZX1114 and 2023ZXZX1084). The funders had no role in study design, data interpretation, manuscript preparation, or the decision to submit.

Data Availability

The deidentified question banks, chatbot response records, scoring datasets, supplementary analysis tables, and analysis code are provided as Multimedia Appendices 3, 4, 7-14. The first-round anonymized scoring dataset used for the primary cross-sectional analyses is provided in Multimedia Appendix 10, and the second-round dataset used for the matched retest sensitivity analysis is provided in Multimedia Appendix 14. Both scoring workbooks contain a README worksheet describing variable definitions, metric ranges, safety coding, score aggregation rules, readability computation, software environment, and data-use conditions. Multimedia Appendix 11 provides the R script for the primary first-round analysis, figure generation, and post hoc matched retest stability analysis; the methods and results of the retest are described in Multimedia Appendix 13. Unless otherwise specified by the publishing journal or repository, the deidentified dataset and analysis code are made available for academic verification, replication, and noncommercial scholarly reuse with appropriate citation. Reuse of chatbot outputs and guideline-derived summaries should comply with the terms of the original chatbot platforms, guideline publishers, and the final journal or repository license. No additional rights are granted for third-party copyrighted material. Additional relevant materials are available from the corresponding author upon reasonable request.

Generative AI Use

ChatGPT (OpenAI) was used during manuscript revision to assist with English-language polishing and code optimization. All AI-assisted wording, interpretive suggestions, and code-related suggestions were manually reviewed and verified against the study data and cited references by the authors before submission. The authors retain full responsibility for the accuracy, integrity, analyses, interpretations, and final content of the manuscript. Generative AI was not used to generate study data, assign expert ratings, conduct independent statistical analyses, or make autonomous clinical judgments.

Authors' Contributions

LF conceived and designed the study, coordinated the rating process, interpreted the findings, and drafted the manuscript. QZ coordinated the development and refinement of the chronic obstructive pulmonary disease (COPD) question bank. J Li and J Lei

XSL•FO
RenderX

organized and curated the question-answer records. ZW performed the statistical analyses and contributed to the preliminary review of the manuscript. RL and XL coordinated the expert rating procedures and quality-control processes. YY critically reviewed the manuscript and provided overall supervision of the study. All authors contributed to the manuscript revision, approved the final version, and agree to be accountable for all aspects of the work.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Study protocol summary.
[DOCX File , 27 KB-Multimedia Appendix 1]

Multimedia Appendix 2

CHART checklist.
[DOCX File , 22 KB-Multimedia Appendix 2]

Multimedia Appendix 3

Chronic obstructive pulmonary disease (COPD) question bank (55 questions).
[DOCX File , 20 KB-Multimedia Appendix 3]

Multimedia Appendix 4

Chronic obstructive pulmonary disease (COPD) 108-to-55 inclusion and exclusion record.
[DOCX File , 44 KB-Multimedia Appendix 4]

Multimedia Appendix 5

Chatbot versions, access type, source status, and query dates.
[DOCX File , 18 KB-Multimedia Appendix 5]

Multimedia Appendix 6

Second-round chatbot versions, access type, source status, and query dates.
[DOCX File , 18 KB-Multimedia Appendix 6]

Multimedia Appendix 7

Question-and-answer (Q&A) record for chronic obstructive pulmonary disease (COPD) large language model (LLM) responses.
[DOCX File , 713 KB-Multimedia Appendix 7]

Multimedia Appendix 8

Chronic obstructive pulmonary disease (COPD) 55-question answer and guideline mapping table.
[DOCX File , 30 KB-Multimedia Appendix 8]

Multimedia Appendix 9

Interrater agreement and pairwise comparison results.
[DOCX File , 67 KB-Multimedia Appendix 9]

Multimedia Appendix 10

Anonymized scoring dataset.
[XLSX File (Microsoft Excel File), 196 KB-Multimedia Appendix 10]

Multimedia Appendix 11

R code for statistical analysis and figure generation.
[TXT File , 91 KB-Multimedia Appendix 11]

XSL•FO
RenderX

Multimedia Appendix 12

Qualitative analysis of unsafe chronic obstructive pulmonary disease (COPD) large language model (LLM) responses.
[DOCX File , 33 KB-Multimedia Appendix 12]

Multimedia Appendix 13

Two-round response stability analysis.
[DOCX File , 30 KB-Multimedia Appendix 13]

Multimedia Appendix 14

Second-round anonymized scoring dataset.
[XLSX File (Microsoft Excel File), 184 KB-Multimedia Appendix 14]

References

1. Global strategy for prevention, diagnosis and management of COPD: 2026 report. Global Initiative for Chronic Obstructive Lung Disease (GOLD). 2026. URL: https://goldcopd.org/2026-gold-report-and-pocket-guide/ [accessed 2026-06-10]
2. Agustí A, Celli BR, Criner GJ, Halpin D, Anzueto A, Barnes P, et al. Global Initiative for Chronic Obstructive Lung Disease 2023 report: GOLD executive summary. Eur Respir J. 2023;61(4):2300239. [FREE Full text] [doi: 10.1183/13993003.00239-2023] [Medline: 36858443]
3. Bourbeau J, Bhutani M, Hernandez P, Aaron SD, Beauchesne M, Kermelly SB, et al. 2023 Canadian Thoracic Society guideline on pharmacotherapy in patients with stable COPD. Chest. 2023;164(5):1159-1183. [doi: 10.1016/j.chest.2023.08.014] [Medline: 37690008]
4. Adeloye D, Song P, Zhu Y, Campbell H, Sheikh A, Rudan I, et al. NIHR RESPIRE Global Respiratory Health Unit. Global, regional, and national prevalence of, and risk factors for, chronic obstructive pulmonary disease (COPD) in 2019: a systematic review and modelling analysis. Lancet Respir Med. 2022;10(5):447-458. [FREE Full text] [doi: 10.1016/S2213-2600(21)00511-7] [Medline: 35279265]
5. GBD Chronic Respiratory Disease Collaborators. Prevalence and attributable health burden of chronic respiratory diseases, 1990-2017: a systematic analysis for the Global Burden of Disease Study 2017. Lancet Respir Med. 2020;8(6):585-596. [FREE Full text] [doi: 10.1016/S2213-2600(20)30105-3] [Medline: 32526187]
6. Wedzicha JA, Miravitlles M, Hurst JR, Calverley PMA, Albert RK, Anzueto A, et al. Management of COPD exacerbations: a European Respiratory Society/American Thoracic Society guideline. Eur Respir J. 2017;49(3):1600791. [FREE Full text] [doi: 10.1183/13993003.00791-2016] [Medline: 28298398]
7. Stolz D, Mkorombindo T, Schumann DM, Agusti A, Ash SY, Bafadhel M, et al. Towards the elimination of chronic obstructive pulmonary disease: a Lancet Commission. Lancet. 2022;400(10356):921-972. [doi: 10.1016/S0140-6736(22)01273-9] [Medline: 36075255]
8. Chronic obstructive pulmonary disease (COPD). World Health Organization. URL: https://www.who.int/news-room/fact-sheets/detail/chronic-obstructive-pulmonary-disease-(copd) [accessed 2026-06-10]
9. Haug CJ, Drazen JM. Artificial intelligence and machine learning in clinical medicine, 2023. N Engl J Med. 2023;388(13):1201-1208. [doi: 10.1056/NEJMra2302038] [Medline: 36988595]
10. Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. N Engl J Med. 2023;388(13):1233-1239. [doi: 10.1056/NEJMsr2214184] [Medline: 36988602]
11. Ayers JW, Poliak A, Dredze M, Leas EC, Zhu Z, Kelley JB, 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):589-596. [FREE Full text] [doi: 10.1001/jamainternmed.2023.1838] [Medline: 37115527]
12. Pan A, Musheyev D, Bockelman D, Loeb S, Kabarriti AE. Assessment of artificial intelligence chatbot responses to top searched queries about cancer. JAMA Oncol. 2023;9(10):1437-1440. [doi: 10.1001/jamaoncol.2023.2947] [Medline: 37615960]
13. Johnson D, Goodman R, Patrinely J, Stone C, Zimmerman E, Donald R, et al. Assessing the accuracy and reliability of AI-generated medical responses: an evaluation of the Chat-GPT model. Res Sq. 2023:rs-2566942. [FREE Full text] [doi: 10.21203/rs.3.rs-2566942/v1] [Medline: 36909565]
14. Lahat A, Shachar E, Avidan B, Glicksberg B, Klang E. Evaluating the utility of a large language model in answering common patients' gastrointestinal health-related questions: are we there yet? Diagnostics (Basel). 2023;13(11):1950. [FREE Full text] [doi: 10.3390/diagnostics13111950] [Medline: 37296802]
15. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nat Med. 2023;29(8):1930-1940. [doi: 10.1038/s41591-023-02448-8] [Medline: 37460753]
16. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. Nature. 2023;620(7972):172-180. [FREE Full text] [doi: 10.1038/s41586-023-06291-2] [Medline: 37438534]

XSL•FO
RenderX

17. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health. 2023;2(2):e0000198. [FREE Full text] [doi: 10.1371/journal.pdig.0000198] [Medline: 36812645]
18. Cascella M, Montomoli J, Bellini V, Bignami E. Evaluating the feasibility of ChatGPT in healthcare: an analysis of multiple clinical and research scenarios. J Med Syst. 2023;47(1):33. [FREE Full text] [doi: 10.1007/s10916-023-01925-4] [Medline: 36869927]
19. Wilhelm TI, Roos J, Kaczmarczyk R. Large language models for therapy recommendations across 3 clinical specialties: comparative study. J Med Internet Res. 2023;25:e49324. [FREE Full text] [doi: 10.2196/49324] [Medline: 37902826]
20. Park YJ, Pillai A, Deng J, Guo E, Gupta M, Paget M, et al. Assessing the research landscape and clinical utility of large language models: a scoping review. BMC Med Inform Decis Mak. 2024;24(1):72. [FREE Full text] [doi: 10.1186/s12911-024-02459-6] [Medline: 38475802]
21. Aydin S, Karabacak M, Vlachos V, Margetis K. Large language models in patient education: a scoping review of applications in medicine. Front Med (Lausanne). 2024;11:1477898. [FREE Full text] [doi: 10.3389/fmed.2024.1477898] [Medline: 39534227]
22. Shool S, Adimi S, Saboori Amleshi R, Bitaraf E, Golpira R, Tara M. A systematic review of large language model (LLM) evaluations in clinical medicine. BMC Med Inform Decis Mak. 2025;25(1):117. [FREE Full text] [doi: 10.1186/s12911-025-02954-4] [Medline: 40055694]
23. Huo B, Boyle A, Marfo N, Tangamornsuksan W, Steen JP, McKechnie T, et al. Large language models for chatbot health advice studies: a systematic review. JAMA Netw Open. 2025;8(2):e2457879. [FREE Full text] [doi: 10.1001/jamanetworkopen.2024.57879] [Medline: 39903463]
24. Wang L, Li J, Zhuang B, Huang S, Fang M, Wang C, et al. Accuracy of large language models when answering clinical research questions: systematic review and network meta-analysis. J Med Internet Res. 2025;27:e64486. [FREE Full text] [doi: 10.2196/64486] [Medline: 40305085]
25. Nigro M, Behring GE, Aliverti A, Angelucci A, Simonds AK, Anzueto A, et al. Accuracy, comprehensiveness and understandability of AI-generated answers to questions from people with COPD: the AIR-COPD Study. Respir Res. 2025;27(1):19. [FREE Full text] [doi: 10.1186/s12931-025-03438-9] [Medline: 41402777]
26. Merç P, Pirinççi C, Cihan E. Evaluation of AI chatbots for patient education and information on chronic obstructive pulmonary disease. Heart Lung. 2026;75:21-25. [doi: 10.1016/j.hrtlng.2025.09.002] [Medline: 40939398]
27. Wu Z, Hu B, Mao Q, Han Y, Zhao Q. Accuracy, reliability, readability, and European Respiratory Society guideline consistency of six generative artificial intelligence chatbots in providing health advice for chronic cough: a cross-sectional comparative assessment. Digit Health. 2026;12:20552076261444223. [FREE Full text] [doi: 10.1177/20552076261444223] [Medline: 42004471]
28. Chen Z, Jiang Y, Jiang H, Xie H, Wu L, Shen Y. Comparative performance of ChatGPT-4.0, DeepSeek, Gemini and Perplexity in answering common questions from patients with COPD. Sci Rep. 2026;16. [doi: 10.1038/s41598-026-70606-2]
29. Poureslami I, FitzGerald JM, Tregobov N, Goldstein RS, Lougheed MD, Gupta S. Health literacy in asthma and chronic obstructive pulmonary disease (COPD) care: a narrative review and future directions. Respir Res. 2022;23(1):361. [FREE Full text] [doi: 10.1186/s12931-022-02290-5] [Medline: 36529734]
30. Jeganathan C, Hosseinzadeh H. The role of health literacy on the self-management of chronic obstructive pulmonary disease: a systematic review. COPD. 2020;17(3):318-325. [FREE Full text] [doi: 10.1080/15412555.2020.1772739] [Medline: 32506965]
31. Shao Y, Hu H, Liang Y, Hong Y, Yu Y, Liu C, et al. Health literacy interventions among patients with chronic diseases: a meta-analysis of randomized controlled trials. Patient Educ Couns. 2023;114:107829. [doi: 10.1016/j.pec.2023.107829] [Medline: 37270933]
32. 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(2):105-111. [FREE Full text] [doi: 10.1136/jech.53.2.105] [Medline: 10396471]
33. 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):165-175. [FREE Full text] [doi: 10.1111/j.1369-7625.2004.00273.x] [Medline: 15117391]
34. 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(15):1244-1245. [doi: 10.1001/jama.1997.03540390074039] [Medline: 9103351]
35. 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(9):2070-2077. [doi: 10.1111/j.1572-0241.2007.01325.x] [Medline: 17511753]
36. CHART Collaborative. Reporting guideline for chatbot health advice studies: the Chatbot Assessment Reporting Tool (CHART) statement. BMJ Med. 2025;4(1):e001632. [FREE Full text] [doi: 10.1136/bmjmed-2025-001632] [Medline: 40761518]
37. Smith EA, Senter RJ. Automated readability index. AMRL TR. 1967:1-14. [Medline: 5302480]

XSL•FO
RenderX

38. Coleman M, Liau TL. A computer readability formula designed for machine scoring. J Appl Psychol. 1975;60(2):283-284. [doi: 10.1037/h0076540]
39. Flesch R. A new readability yardstick. J Appl Psychol. 1948;32(3):221-233. [doi: 10.1037/h0057532] [Medline: 18867058]
40. Kincaid J, Fishburne RJ, Rogers R, Chissom B. Derivation of new readability formulas (Automated Readability Index, Fog Count, and Flesch Reading Ease Formula) for navy enlisted personnel. Research Branch Report 8-75. 1975. URL: https://apps.dtic.mil/sti/pdfs/ADA006655.pdf [accessed 2026-06-12]
41. Gunning R. The Technique of Clear Writing. New York, United States. McGraw-Hill; 1952.
42. McLaughlin GH. SMOG grading: a new readability formula. J Read. 1969;12(8):639-646. [FREE Full text]
43. Benoit K, Watanabe K, Wang H, Lua JW, Kuha J. quanteda.textstats: textual statistics for the quantitative analysis of textual data. R package documentation. URL: https://quanteda.io/reference/textstat_readability.html [accessed 2026-06-12]

Abbreviations

ARI: Automated Readability Index
CHART: Chatbot Assessment Reporting Tool
CLI: Coleman-Liau Index
COPD: chronic obstructive pulmonary disease
EQIP: Ensuring Quality Information for Patients
FKGL: Flesch-Kincaid Grade Level
FRES: Flesch Reading Ease Score
GFI: Gunning Fog Index
GOLD: Global Initiative for Chronic Obstructive Lung Disease
GQS: Global Quality Score
ICC: intraclass correlation coefficient
JAMA: Journal of the American Medical Association
LLM: large language model
SMOG: Simple Measure of Gobbledygook

Edited by I Steenstra; submitted 18.Jun.2026; peer-reviewed by X Zhan, F Pu; comments to author 25.Aug.2026; revised version received 19.Oct.2026; accepted 20.Oct.2026; published 09.Nov.2026

Please cite as:
Fan L, Liu R, Wu Z, Li J, Lei J, Zhang Q, Li X, Yang Y
Safety, Accuracy, Empathy, Information Quality, and Readability of Chronic Obstructive Pulmonary Disease Health Advice Generated by 6 Large Language Models: Cross-Sectional Comparative Evaluation Study
JMIR Med Inform 2026;14:e104836
URL: https://medinform.jmir.org/2026/1/e104836
doi: 10.2196/104836
PMID:

©Liping Fan, Ru Liu, Zhishui Wu, Jingjing Li, Jiao Lei, Qingqing Zhang, Xinghua Li, Yanming Yang. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 09.Nov.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX