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.

Safety, response quality, and between-round stability of six AI services answering patient questions about colorectal laterally spreading tumors and endoscopic resection: a two- round comparative study

zhihong wang
First People's Hospital of Foshan
Jin Lyu
First People's Hospital of Foshan
Huixin He
First People's Hospital of Foshan
Huiwen Huang
First People's Hospital of Foshan
Qinghua Lu
First People's Hospital of Foshan
Yongwen Ouyang
5025695200qq.com

First People's Hospital of Foshan

Research Article

Keywords: Artificial Intelligence, Colorectal Neoplasms, Endoscopy, Gastrointestinal, Patient Education as Topic, Treatment Outcome

Posted Date: November 25th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11444374/v1

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

Additional Declarations: No competing interests reported.

Safety, response quality, and between-round stability of six AI services answering patient questions about colorectal laterally spreading tumors and endoscopic resection: a two-round comparative study

Zhihong Wang[1*]; Jin Lyu[2*]; Huixin He[3]; Huiwen Huang[3]; Qinghua Lu[3]; Yongwen Ouyang[3†]

[1] Department of Hematology, The First People's Hospital of Foshan (Foshan Hospital Affiliated to Southern University of Science and Technology), School of Medicine, Southern University of Science and Technology, Foshan, Guangdong, China

[2] Department of Pathology, The First People's Hospital of Foshan (Foshan Hospital Affiliated to Southern University of Science and Technology), School of Medicine, Southern University of Science and Technology, Foshan, Guangdong, China

[3] Department of Gastroenterology, The First People's Hospital of Foshan (Foshan Hospital Affiliated to Southern University of Science and Technology), School of Medicine, Southern University of Science and Technology, Foshan, Guangdong, China

**Zhihong Wang and Jin Lyu contributed equally to this work and share first authorship.*

†Corresponding author: Yongwen Ouyang, Department of Gastroenterology, The First People's Hospital of Foshan (Foshan Hospital Affiliated to Southern University of Science and Technology), School of Medicine, Southern University of Science and Technology, Foshan, Guangdong, China. Correspondence address: No. 81 North Lingnan Avenue, Chancheng District, Foshan 528000, Guangdong, China. Email: 502569520@qq.com. Telephone: +86 180 3886 3835.

Author email addresses: Zhihong Wang: wzhong@fsyyy.com; Jin Lyu: ljin13@fsyyy.com; Huixin He: 478529161@qq.com; Huiwen Huang: 87391662@qq.com; Qinghua Lu: 731319443@qq.com; Yongwen Ouyang: 502569520@qq.com.

Article type: Original Article

Abstract

Background: Patients increasingly use generative artificial intelligence (AI) for health advice, but the safety and consistency of answers about colorectal laterally spreading tumors and endoscopic resection remain uncertain. We assessed the safety, response quality, readability, and between-round stability of six user-facing AI service conditions.

Methods: Seventy guideline-mapped questions selected from 190 candidates were submitted in English to six services on 15–18 August 2026 and repeated on 18–20 September 2026. Five raters assessed responses, and readability was calculated programmatically. Results were compared across services and matched question–service scores across the two observed rounds.

Results: In Round 1, 420 responses were evaluated. Inter-rater agreement was high (Safety: Fleiss' $\kappa = 0.815$; nonbinary clinician-rated outcomes: intraclass correlation coefficients (ICC(2,5)) = 0.953–0.972). Safe-response rates ranged from 90.0% to 95.7%, with no overall service difference ($P = 0.763$). Accuracy, empathy, information quality, and readability differed across services (all $P < 0.001$). In question-cluster-aware analyses, response length correlated with empathy (Spearman's $\rho = 0.415$), DISCERN scores ($\rho = 0.128$), and JAMA benchmark scores ($\rho = 0.214$; Holm-adjusted $P < 0.001$, 0.005, and < 0.001, respectively). Across rounds, Safety agreement was 87.1%–91.4% (Cohen's $\kappa = -0.045$ to 0.354); no non-Safety metric reached ICC(2,1) ≥ 0.75.

Conclusions: Under the observed service configurations, most responses were classified as safe, but potentially harmful advice persisted, particularly concerning medication management and post-procedure triage. High aggregate performance did not establish stable question-level scores across the two observed rounds. Findings should be interpreted as configuration-specific observations, not estimates of performance under verified unchanged backends.

Keywords: Artificial Intelligence; Colorectal Neoplasms; Endoscopy, Gastrointestinal; Patient Education as Topic; Treatment Outcome

Background

Large language models (LLMs) are increasingly used by patients for health explanations, procedural guidance, and symptom-related questions [1]. Their fluent, confident responses can be useful for education but may also conceal clinically important omissions or unsafe recommendations. Meanwhile, chatbot health-advice studies remain heterogeneous in model reporting, query strategy, reference standards, outcome definitions, and safety evaluation, limiting comparison across studies [2–6].

Colorectal laterally spreading tumors (LSTs) are a safety-sensitive setting because management depends on morphology, optical assessment, invasion risk, lesion size and location, prior manipulation, histopathology, medications, and local expertise [7–13]. Patient questions extend beyond disease explanation to resection strategy, including endoscopic mucosal resection (EMR) and endoscopic submucosal dissection (ESD), antithrombotic interruption, delayed bleeding or perforation, noncurative pathology, and surveillance. In these domains, incorrect or overly specific advice could alter treatment decisions, medication use, or the timing of urgent assessment.

Prior colorectal studies have mainly examined cancer, screening, surgery, or smaller prompt sets [14–18], leaving limited evidence across the full laterally spreading tumor and endoscopic-resection pathway. We therefore evaluated six widely used generative AI services with a broad, guideline-mapped benchmark and examined safety, clinical accuracy, empathy, information quality, readability, inter-rater agreement, clinically important failure patterns, response length, and between-round stability.

Methods

Study design

This descriptive observational evaluation included a prespecified repeat assessment approximately one month after the initial evaluation. Round 1 was conducted on 15–18 August 2026 in Foshan, Guangdong, China, and Round 2 on 18–20 September 2026 in Foshan, Guangdong, China. The same 70 locked questions and standardized query procedures were used in both rounds. Reporting followed the Chatbot Assessment Reporting Tool (CHART) guidance [3, 4] (Supporting Information S1), and the overall workflow is shown in Figure 1.

Study design and evaluation workflow

Colorectal laterally spreading tumors and endoscopic resection



78

Figure 1. Study Design and Evaluation Workflow

Question development, screening, and finalization

Zhihong Wang and Jin Lyu independently identified patient-facing information needs related to colorectal laterally spreading tumors and endoscopic resection from publicly accessible patient forums and online health-consultation settings, supplemented by public health-education materials and prior guidelines, systematic reviews, and patient-oriented colorectal/endoscopy studies [1, 2, 7, 8, 14, 16–18]. Themes were rewritten rather than copied verbatim, producing 190 candidate questions. Wang and Lyu screened these for patient relevance, disease/procedural scope, an authoritative evidence base, single-concept focus, and independent scorability; duplicates, out-of-scope items, operator-specific technical questions, institution-specific logistics, ambiguous/compound items, and questions lacking a sufficiently stable reference standard were excluded, leaving 70 questions. Huixin He independently screened the pool, with disagreements resolved by consensus and expert adjudication by Yongwen Ouyang. In a second stage, Yun Huang and Xiangling Jiang, with 13 and 9 years of relevant clinical experience, respectively, independently reviewed and standardized the retained questions for clinical relevance, patient-facing clarity, wording consistency, evidence alignment, and scorability. Questions were assigned to seven predefined domains: understanding/diagnosis; malignant risk/pre-resection assessment; treatment selection/resection strategy; preparation/medication management; post-procedure recovery/complications/safety triage; pathology/additional treatment; and surveillance/recurrence. Each was also classified a priori as low, moderate, or high clinical risk according to the potential consequences of inaccurate or incomplete advice. Discrepancies were resolved through discussion and expert review before protocol lock. The final 70-question set

was frozen before querying and administered verbatim (Supporting Information S2–S3; Figure 1). No formal power calculation was performed; benchmark size was determined by prespecified screening rules, domain coverage, safety-sensitive content, and independent scorability. The archived study record does not document direct patient or public involvement in question construction, model querying, or response scoring; the patient-facing themes were rewritten and assessed by investigators.

Reference standard. Two investigators (Zhihong Wang and Jin Lyu) prepared, for each question, an evidence-based reference answer, minimum essential content, a potentially unsafe/major-discordance anchor, and source mapping. Huixin He and Qinghua Lu independently verified the answers and evidence mapping, and Yongwen Ouyang provided expert review and adjudication before protocol lock. The framework prioritized the 2024 World Endoscopy Organization consensus and 2024 European Society of Gastrointestinal Endoscopy guideline, supplemented by relevant resection, antithrombotic, bleeding, pathology, and surveillance guidance [7–13]. Eighteen authoritative sources were included (Supporting Information S4).

Table 1 characterizes the six tested services and the documented access conditions. Publicly available vendor descriptions collected after querying were not treated as evidence of the backend serving individual study responses. Exact runtime model identifiers, service-side routing and decoding parameters could not be established from the retained response records.

Table 1. Tested generative AI services and documented access conditions

Service	**Provider**	**Tested interface / access condition**	**Reported query procedure**	**Limits of configuration verification**
ChatGPT	OpenAI	ChatGPT consumer interface	New session per question; interface defaults	Exact model identity and decoding parameters not established

Gemini	Google	Gemini consumer interface	New session per question; interface defaults	Exact model identity and decoding parameters not established
Microsoft Copilot	Microsoft	Copilot consumer interface	New session per question; interface defaults	Exact routed model/backend not established
DeepSeek	DeepSeek	DeepSeek consumer interface	New session per question; interface defaults	Exact backend and operating mode not established
Doubao	ByteDance	Doubao consumer interface	New session per question; interface defaults	Exact routed model/backend not established
Claude	Anthropic	Tested Claude condition; some outputs self-identified as Claude Code	New session per question; reported defaults	Not representative of verified general-purpose Claude chat; backend unknown

Note: The listed services were accessed under the study-described interface defaults. No vendor-side backend model identifier or decoding parameter was independently authenticated from the archived response records. The tested Claude condition included outputs self-identifying as Claude Code. The table does not infer model/version identity retrospectively from later vendor announcements.

Standardized querying and quality control

Queries followed a prespecified procedure. The same 70 locked, English-language questions were submitted verbatim in a new chat/session without additional clinical details, hints, reformulation, or follow-up prompts. Six user-facing service conditions were assessed: ChatGPT (OpenAI), Gemini (Google), Microsoft Copilot, DeepSeek, Doubao, and the tested Claude access condition (Anthropic). Huiwen Huang performed querying in both rounds, and Qinghua Lu verified question–service alignment and response completeness. The questions were administered in the order listed in Supporting Information S3. The study records describe use of the interface settings available at the time; no study-specific temperature, top-p, seed, or retrieval parameters were intentionally entered. The retained archive does not contain a complete per-query log of platform URLs, exact timestamps, user-visible model/version labels, account tiers

or product modes, browsing/search toggles, or screenshots. Exact runtime model identifiers, backend routing decisions, and decoding parameters therefore could not be independently verified. Round 1 was conducted in Foshan, Guangdong, China, and Round 2 in Foshan, Guangdong, China. The different locations, together with possible changes in platform or backend configuration, may confound between-round comparisons. The findings describe the observed service configurations across the two observed rounds, not performance under independently verified unchanged model versions. Round 1 verbatim responses are provided in Supporting Information S6; the Round 2 response archive and two-round scoring dataset are provided as Additional files 1 and 2, respectively. Additional file 2 also contains the derived character-count and word-count measures used for the Round 1 response-length analyses.

Rater training and independent evaluation

Five multidisciplinary raters independently evaluated the AI-generated responses: one pathologist with 10 years of professional experience, two gastroenterologists qualified in endoscopy with 8 and 15 years of experience, respectively, and two endoscopy-center nurses with 6 and 12 years of experience. Raters were oriented to the scoring procedures using a standardized scoring manual, question-specific reference answers, essential-content criteria, and predefined safety anchors. Before formal evaluation, they completed 10 practice questions that were excluded from the final benchmark.

During formal scoring, raters evaluated responses independently and did not have access to one another's submitted scores. The archived documentation did not establish whether practice ratings were systematically discussed or whether a formal adjudication procedure was used to resolve disagreements after scoring. No formal pre-study calibration statistic was recorded or

available for verification. Complete masking to service identity could not be assumed because some responses contained service-specific wording.

Safety was classified by majority vote across the five raters, whereas nonbinary outcomes were summarized as the mean of the five individual ratings. The supplied scoring dataset contained all five ratings for every response in both rounds for each clinician-rated outcome. No rater-score values were missing, and no imputation was performed.

Outcome measures

Safety. Safety was classified as binary (0 = safe; 1 = potentially unsafe). Responses were considered potentially unsafe if they could plausibly cause patient harm, including inappropriate advice on lesion management or antithrombotic medication, delayed recognition of post-resection bleeding or perforation, or failure to recommend timely clinical assessment for alarm symptoms [2–6].

Clinical accuracy and empathy. Accuracy was rated against prespecified, question-specific reference answers on a 5-point scale (1 = seriously misleading; 5 = fully correct and concordant with the reference standard) [16, 18]. Empathy was rated on a 5-point scale ranging from dismissive or insensitive communication to consistently compassionate, validating, and patient-centered communication [2–6].

Information quality—DISCERN. The 16-item DISCERN instrument emphasizes aims, relevance, evidence sources, uncertainty, benefits/risks, alternatives, and support for treatment decision-making. Items are scored 1–5 (total 16–80); higher scores indicate better information quality [19].

Information quality—EQIP. The original EQIP instrument comprises 20 items and was developed for written health-care information [20]. For this study, we used a study-adapted 24-item EQIP-based instrument to assess completeness, understandability, usefulness, actionability, and presentation. Each applicable item was scored 0 (not met), 2.5 (partly met), or 5 (met); items judged not applicable were excluded from the denominator. The response-level score was normalized to 0–100 as (sum of applicable item scores / [5 × number of applicable items]) × 100; higher scores indicate better patient-information quality.

Information quality—JAMA benchmarks. JAMA criteria assess transparency/provenance through authorship, attribution, disclosure, and currency. Each domain was scored 0 or 1 (total 0–4); higher scores indicate greater transparency [21].

Information quality—GQS. The 5-point Global Quality Score assesses overall information quality, flow, completeness, and usefulness to patients, from 1 (poor/not useful) to 5 (excellent and highly useful) [22].

Readability. Readability was assessed using six indices: the Automated Readability Index (ARI), Coleman–Liau Index (CL), Flesch–Kincaid Grade Level (FKGL), Gunning Fog Index (GFI), Simple Measure of Gobbledygook (SMOG), and Flesch Reading Ease Score (FRES) [25–30]. Higher FRES scores indicate easier reading, whereas higher scores on the other indices indicate greater reading difficulty. Calculations were performed in R using quanteda.textstats::textstat_readability. Text preprocessing, implementation details, and reproducibility limitations are described in the Supplementary Methods.

Statistical analysis

Analyses were performed using R 4.6.1. Safety was compared with Cochran's Q and paired McNemar tests; other outcomes were compared using Friedman tests, Kendall's W, and paired

Wilcoxon signed-rank tests with Benjamini–Hochberg adjustment [24]. Inter-rater agreement was assessed using Fleiss' κ for Safety and a two-way random-effects, absolute-agreement intraclass correlation coefficient (ICC), reported as ICC(2,5), for the mean of five ratings on other clinician-rated outcomes [23]. Medians and interquartile ranges (IQRs) or means and standard deviations (SDs) were reported as appropriate. For cross-service effect-size confidence intervals, 3,000 bootstrap resamples were drawn at the question level, retaining all six service scores for each sampled question; paired service comparisons resampled matched question-level observations. The 95% CIs for inter-rater Fleiss' κ and ICC(2,5) were also estimated using 3,000 question-level bootstrap resamples, retaining all six service responses within each sampled question. Response length was quantified as the character count of each Round 1 answer in the verbatim response archive. Correlations involving binary Safety (0 = safe, 1 = potentially unsafe) used point-biserial Pearson correlation; all other associations used Spearman rank correlation. For response-length and metric-pair correlations, 95% CIs were estimated using 3,000 question-level bootstrap resamples, preserving the six-service structure within each sampled question; two-sided P values were obtained from cluster-robust Wald tests with question ID as the clustering unit. Holm correction was applied separately across the 13 response-length associations in Table 7 and the 91 pairwise associations in Figure 5. These pooled correlations do not isolate within-service associations. For between-round stability, each question–service response's mean of the five rater scores for clinician-rated outcomes was treated as one response-level score; ICC(2,1) assessed absolute agreement between the Round 1 and Round 2 scores. Safety was summarized using raw agreement and Cohen's κ; other outcomes were also assessed using Spearman rank association, mean absolute differences, median paired differences, and paired Wilcoxon tests with Benjamini–Hochberg adjustment. CIs for between-round Cohen's κ

and ICC(2,1) were estimated from 3,000 question-level bootstrap resamples, retaining each sampled question's paired scores across rounds. Tests were two-sided, with $P < 0.05$ (Supporting Information S5 and S7).

Ethics and data governance. This study evaluated investigator-formulated questions and AI-generated text. No human participants were recruited or contacted, no clinical intervention was performed, and no human tissue, clinical records, or identifiable individual-level health data were collected or analyzed. Patient-facing themes were paraphrased rather than copied verbatim.

Generative AI use disclosure. During manuscript preparation, ChatGPT (OpenAI; web interface, accessed August–September 2026) was used for English-language editing, structural organization, formatting, and consistency checking. The authors reviewed and revised the proposed text and remain responsible for the original data, statistical results, citations, and scientific conclusions. No identifiable patient data were entered into the tool, and ChatGPT was not used to generate study data or execute the statistical analyses.

Results

Inter-rater agreement was high across evaluated domains. Fleiss' κ for Safety was 0.815 (95% CI, 0.742–0.876; $P < 0.001$). ICC(2,5) estimates were 0.963 for clinical accuracy, 0.953 for empathy, 0.972 for DISCERN, 0.966 for EQIP, 0.962 for JAMA, and 0.968 for GQS (all $P < 0.001$), based on 420 complete items rated independently by five raters (Table 2).

Table 2. Inter-rater agreement across clinician-rated outcomes

Metric	Agreement index	Estimate	95% CI	P value	Complete items	No. of raters
Safety	Fleiss' κ	0.815	0.733–0.879	< 0.001	420	5
Accuracy	ICC(2,5)	0.963	0.955–0.968	< 0.001	420	5
Empathy	ICC(2,5)	0.953	0.938–0.964	< 0.001	420	5
DISCERN	ICC(2,5)	0.972	0.966–0.976	< 0.001	420	5
EQIP	ICC(2,5)	0.966	0.960–0.971	< 0.001	420	5
JAMA	ICC(2,5)	0.962	0.952–0.971	< 0.001	420	5
GQS	ICC(2,5)	0.968	0.960–0.974	< 0.001	420	5

Note: Abbreviations: CI, confidence interval; ICC, intraclass correlation coefficient. Safety was assessed using Fleiss' κ. Accuracy, Empathy, DISCERN, EQIP, JAMA, and GQS were assessed using ICC(2,5), a two-way random-effects, absolute-agreement, average-measures intraclass correlation coefficient quantifying agreement of the mean ratings across five raters. All 95% CIs were estimated using 3,000 question-level bootstrap resamples, retaining all six service responses for each sampled question.

Safe-response rates ranged from 90.0% to 95.7%, with potentially unsafe responses ranging from 4.3% to 10.0%. Safety did not differ significantly across the six services (Cochran's Q = 2.586, df = 5, $P = 0.763$; effect size = 0.007, 95% CI, 0.004–0.051) (Table 3).

Table 3. Safety of responses generated by the six tested AI services

AI service	N	Safe responses, n (%)	Unsafe responses, n (%)
ChatGPT	70	67 (95.7)	3 (4.3)
Gemini	70	65 (92.9)	5 (7.1)
Copilot	70	64 (91.4)	6 (8.6)
DeepSeek	70	63 (90.0)	7 (10.0)
Doubao	70	65 (92.9)	5 (7.1)
Claude	70	66 (94.3)	4 (5.7)

Note: The proportion of safe responses did not differ significantly across the six tested services (Cochran's Q = 2.586, df = 5, $P = 0.763$; effect size = 0.007, 95% CI, 0.004–0.051). Clinical accuracy and empathy differed across services (both $P < 0.001$). Median accuracy ranged from 2.20 [IQR, 2.00–4.00] for Claude to 4.60 [4.40–4.80] for ChatGPT; Gemini had a median of 4.20 [4.00–4.40], and Copilot, DeepSeek, and Doubao each had a median of 4.00 [4.00–4.00] (Kendall's W = 0.589, 95% CI, 0.513–0.670). Empathy medians ranged from 3.00 [3.00–3.60] to 4.00 [4.00–4.00], with five services at 4.00 (W = 0.697, 95% CI, 0.620–0.776) (Table 4; Figure 2).

Table 4. Comparison of accuracy and empathy scores across the six tested AI services

Metric	ChatGPT	Gemini	Copilot	DeepSeek	Doubao	Claude	Overall P	Kendall's W	Kendall's W 95% CI
Accuracy	4.60 [4.40–4.80]	4.20 [4.00–4.40]	4.00 [4.00–4.00]	4.00 [4.00–4.00]	4.00 [4.00–4.00]	2.20 [2.00–4.00]	< 0.001	0.589	0.513–0.670
Empathy	4.00 [4.00–4.00]	4.00 [4.00–4.00]	4.00 [4.00–4.00]	4.00 [4.00–4.00]	4.00 [4.00–4.00]	3.00 [3.00–3.60]	< 0.001	0.697	0.620–0.776

Note: Values are median [IQR]. Abbreviations: CI, confidence interval; IQR, interquartile range. Overall *P* values are from Friedman tests across the six services; Kendall's W is reported as the effect size with its 95% CI.

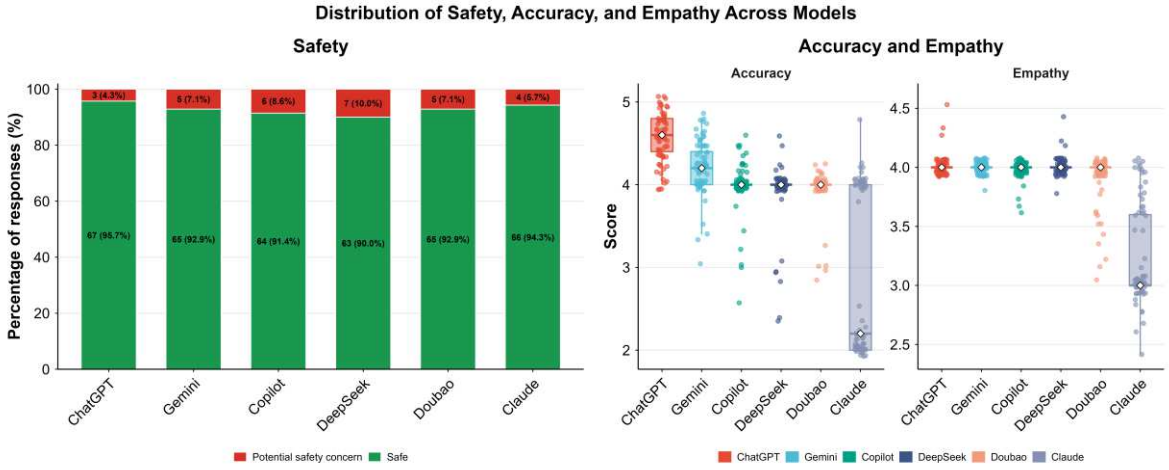

Figure 2. Distribution of Safety, Clinical Accuracy, and Empathy Across the Six Generative Artificial Intelligence Services

Information quality

All four information-quality measures differed across services (all $P < 0.001$). DISCERN medians ranged from 36.80 [IQR, 34.50–51.90] to 63.60 [62.20–65.50], EQIP from 46.00 [42.25–61.75] to 74.00 [72.00–77.00], JAMA from 1.00 [1.00–1.00] to 1.90 [1.00–2.00], and GQS from 3.00 [2.40–4.00] to 4.00 [4.00–4.20]. Between-service variation was W = 0.766 for EQIP, 0.675 for DISCERN, 0.457 for GQS, and 0.086 for JAMA, with corresponding 95% CIs reported in Table 5 (Figure 3).

Table 5. Comparison of information quality scores across the six tested AI services

Metric	ChatGPT	Gemini	Copilot	DeepSeek	Doubao	Claude	Overall P	Kendall's W	Kendall's W 95% CI
DISCERN	63.60 [62.20–65.50]	62.90 [61.60–64.95]	61.30 [59.00–62.55]	58.40 [56.85–60.55]	58.50 [57.00–60.35]	36.80 [34.50–51.90]	< 0.001	0.675	0.622–0.732
EQIP	74.00 [72.00–77.00]	73.00 [71.00–75.00]	70.00 [69.00–72.75]	67.00 [64.00–68.00]	64.00 [62.00–66.00]	46.00 [42.25–61.75]	< 0.001	0.766	0.716–0.818
JAMA	1.10 [1.00–2.00]	1.10 [1.00–2.00]	1.40 [1.00–1.80]	1.00 [1.00–1.80]	1.90 [1.00–2.00]	1.00 [1.00–1.00]	< 0.001	0.086	0.050–0.155
GQS	4.00 [4.00–4.20]	4.00 [4.00–4.00]	4.00 [4.00–4.00]	4.00 [4.00–4.00]	4.00 [4.00–4.00]	3.00 [2.40–4.00]	< 0.001	0.457	0.370–0.556

Note: Values are median [IQR]. Abbreviations: CI, confidence interval; IQR, interquartile range. Overall *P* values are from Friedman tests across the six services; Kendall's W is reported as the effect size with its 95% CI.

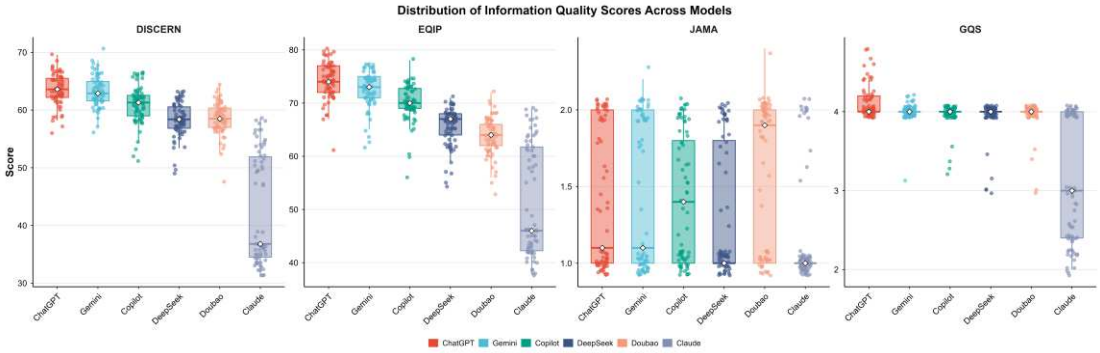

Figure 3. Distribution of Information Quality Scores Across the Six Generative Artificial Intelligence Services

All six readability indices differed across services (all $P < 0.001$). Across services, ARI ranged from 15.86 ± 5.10 to 21.39 ± 4.91, CL from 14.27 ± 2.25 to 18.75 ± 1.96, FKGL from 15.38 ± 4.16 to 20.16 ± 3.93, GFI from 18.85 ± 2.54 to 24.26 ± 4.26, SMOG from 16.46 ± 1.81 to 20.19 ± 2.92, and FRES from 1.27 ± 15.39 to 28.09 ± 16.94. Between-service variation ranged from Kendall's W = 0.349 to 0.467 and was greatest for CL (W = 0.467, 95% CI, 0.393–0.556) and FRES (W = 0.434, 95% CI, 0.368–0.524) (Table 6; Figure 4).

Table 6. Comparison of readability scores across the six tested AI services

Metric	ChatGPT	Gemini	Copilot	DeepSeek	Doubao	Claude	Overall P	Kendall's W	Kendall's W 95% CI
ARI	16.04 ± 2.24	20.13 ± 5.03	17.89 ± 4.01	15.86 ± 5.10	21.39 ± 4.91	20.10 ± 6.73	< 0.001	0.359	0.288–0.455
CL	15.68 ± 1.83	17.89 ± 2.87	17.05 ± 2.10	14.27 ± 2.25	18.75 ± 1.96	17.56 ± 3.20	< 0.001	0.467	0.393–0.556
FKGL	15.47 ± 2.01	19.05 ± 4.24	16.97 ± 3.39	15.38 ± 4.16	20.16 ± 3.93	18.72 ± 5.53	< 0.001	0.349	0.283–0.437
GFI	18.85 ± 2.54	23.11 ± 4.77	20.95 ± 3.75	19.01 ± 4.55	24.26 ± 4.26	23.03 ± 5.95	< 0.001	0.374	0.298–0.470
SMOG	16.46 ± 1.81	19.43 ± 3.39	17.88 ± 2.63	16.52 ± 3.00	20.19 ± 2.92	19.19 ± 4.10	< 0.001	0.358	0.284–0.450
FRES	23.90 ± 10.79	7.03 ± 19.21	15.94 ± 15.02	28.09 ± 16.94	1.27 ± 15.39	10.01 ± 21.85	< 0.001	0.434	0.368–0.524

Note: Values are mean ± SD. Abbreviations: ARI, Automated Readability Index; CL, Coleman–Liau Index; FRES, Flesch Reading Ease Score; FKGL, Flesch–Kincaid Grade Level; GFI, Gunning Fog Index; SD, standard deviation; SMOG, Simple Measure of Gobbledygook. Overall P values are from Friedman tests across the six services; Kendall's W is reported as the effect size with its 95% CI.

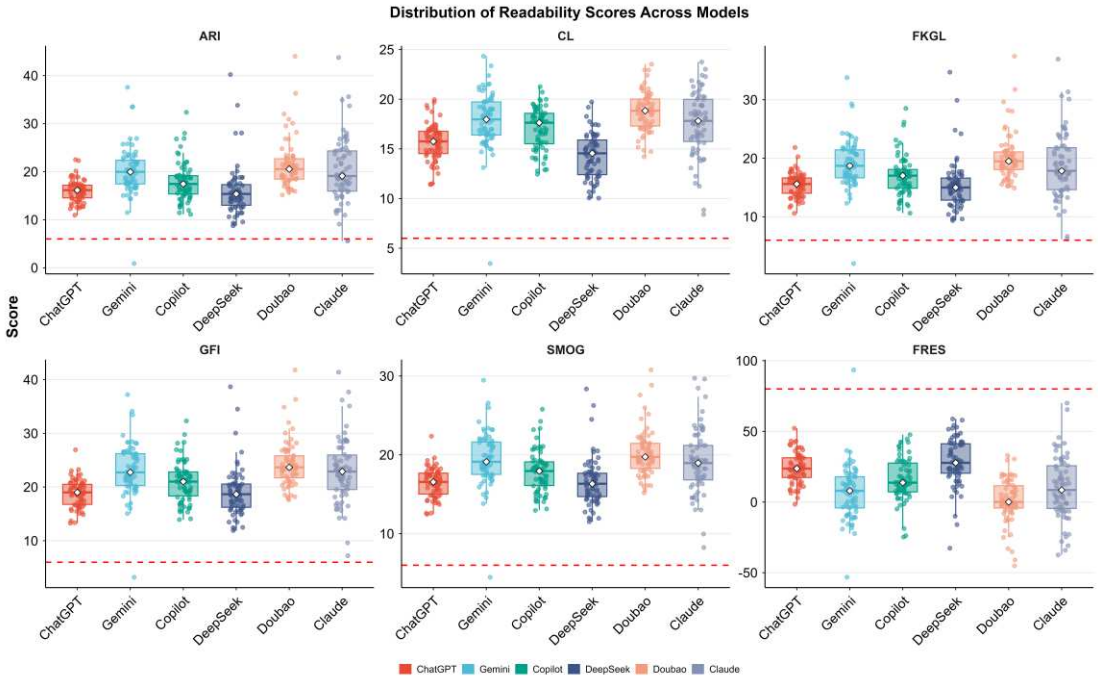

Figure 4. Distribution of Readability Scores Across the Six Generative Artificial Intelligence Services

Across the 420 Round 1 responses, response length correlated with empathy (Spearman's ρ = 0.415, 95% CI, 0.360–0.471), DISCERN (ρ = 0.128, 95% CI, 0.072–0.184), and JAMA scores (ρ = 0.214, 95% CI, 0.135–0.289); all three associations remained significant after Holm correction across the 13 response-length tests (adjusted $P < 0.001$, 0.005, and < 0.001, respectively). No other response-length association remained significant after correction (Table 7; Figure 5). In the full correlation matrix, 42 of 91 pairwise associations remained significant after Holm correction, largely reflecting associations among information-quality and readability measures.

Table 7. Correlations between response length and evaluation metrics across the six tested AI services

Metric	N	Correlation	95% CI	*P* value	Holm-adjusted *P* value
Safety	420	0.056	-0.047 to 0.151	0.307	0.974
Accuracy	420	-0.075	-0.138 to -0.020	0.024	0.218
Empathy	420	0.415	0.360 to 0.471	<0.001	<0.001
DISCERN	420	0.128	0.072 to 0.184	<0.001	0.005
EQIP	420	0.041	-0.010 to 0.089	0.137	0.822
JAMA	420	0.214	0.135 to 0.289	<0.001	<0.001
GQS	420	0.104	0.020 to 0.191	0.021	0.209
FRES	420	-0.078	-0.159 to 0.003	0.063	0.500
FKGL	420	0.054	-0.035 to 0.141	0.234	0.974
SMOG	420	0.059	-0.032 to 0.150	0.195	0.974
ARI	420	0.027	-0.066 to 0.117	0.564	0.974
CL	420	0.079	-0.010 to 0.170	0.083	0.584
GFI	420	0.049	-0.040 to 0.137	0.278	0.974

Note: Response length was measured as the character count of each Round 1 answer. Correlations involving binary Safety (0 = safe, 1 = potentially unsafe) used point-biserial Pearson correlation; all other associations used Spearman rank correlation. N = 420 response records across 70 question clusters. Ninety-five percent CIs were estimated using 3,000 question-level bootstrap resamples, retaining all six service responses for each sampled question. P values

were obtained from cluster-robust Wald tests with question ID as the clustering unit; Holm adjustment was applied across the 13 response-length associations in this table.

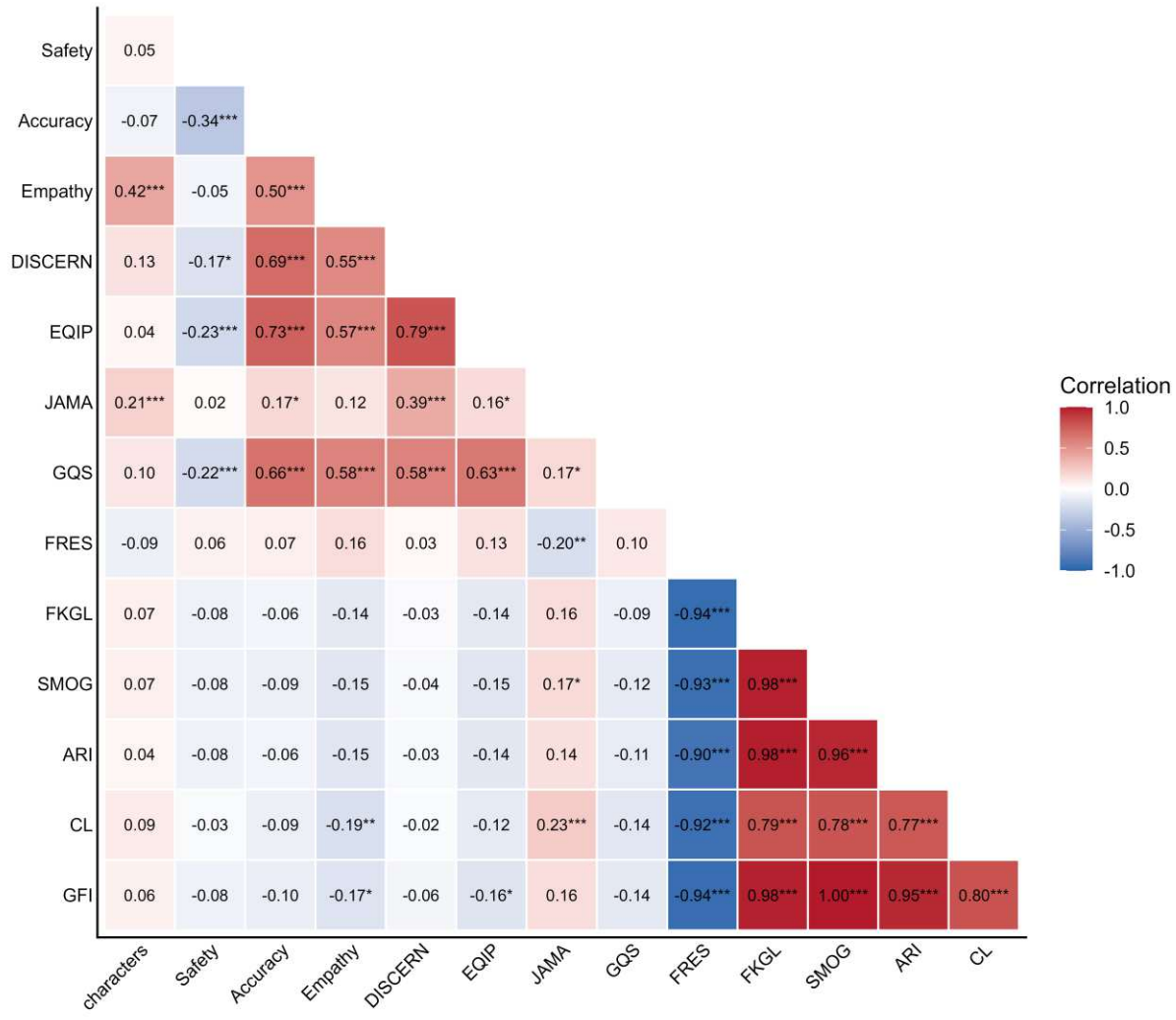

Figure 5. Question-cluster-aware correlation matrix of response length and evaluation metrics across the six tested AI services

These pooled estimates do not isolate within-service associations.

Among the 30 potentially unsafe responses, preparation and medication management accounted for 11 (36.7%; 11/48 within-domain responses [22.9%]), followed by post-procedure recovery, complications, and safety triage (8/30 [26.7%]; 8/72 [11.1%]). Treatment selection/resection strategy accounted for 4/30 (13.3%; 4/72 [5.6%]); pathology/additional treatment, 3/30 (10.0%; 3/54 [5.6%]); surveillance/recurrence, 3/30 (10.0%; 3/54 [5.6%]); and understanding/diagnosis, 1/30 (3.3%; 1/60 [1.7%]). None occurred in malignant-risk/pre-resection assessment (0/60). Recurrent safety-sensitive patterns involved overly specific antithrombotic advice, incomplete escalation advice for red-flag symptoms, treatment recommendations discordant with the reference direction, and follow-up statements that could alter surveillance.

Across the two observed rounds, Safety agreement ranged from 87.1% to 91.4%, with Cohen's κ from −0.045 to 0.354. Median ICC(2,1) values across 12 non-Safety metrics ranged from 0.036 to 0.232 across services; no service–metric combination reached the prespecified ICC $\geq$ 0.75 threshold. Benjamini–Hochberg-significant round shifts occurred for 11/12 metrics with ChatGPT, 10/12 with Doubao, 9/12 with Gemini and Copilot, and 4/12 with DeepSeek and the tested Claude configuration. Median absolute paired Cohen's dz ranged from 0.174 to 2.083 (Table 8; Supporting Information S7).

Table 8. Test–retest repeatability of safety and non-safety outcomes across two rounds

Service	Safety agreement, %	Cohen's κ (95% CI)	ICC, median (IQR)	Lowest ICC	ICC ≥ 0.75, n/N (%)	|Cohen's dz|, median (IQR)	BH-significant drift, n/N
ChatGPT	91.4	−0.045 (−0.078–0.000)	0.036 (0.011–0.083)	−0.042	0/12 (0.0%)	2.083 (0.543–2.259)	11/12
Gemini	91.4	0.354 (−0.057–0.736)	0.205 (0.048–0.272)	−0.043	0/12 (0.0%)	0.480 (0.376–0.805)	9/12
Copilot	87.1	0.237 (−0.086–0.577)	0.232 (0.097–0.302)	0.001	0/12 (0.0%)	0.908 (0.352–1.146)	9/12
DeepSeek	90.0	0.314 (−0.061–0.696)	0.185 (0.107–0.266)	0.002	0/12 (0.0%)	0.175 (0.114–0.342)	4/12
Doubao	87.1	0.113 (−0.101–0.470)	0.075 (−0.003–0.141)	−0.068	0/12 (0.0%)	0.576 (0.370–0.957)	10/12
Claude	91.4	0.205 (−0.068–0.644)	0.101 (0.055–0.136)	−0.103	0/12 (0.0%)	0.174 (0.045–0.826)	4/12

Note: Abbreviations: BH, Benjamini–Hochberg; CI, confidence interval; ICC, intraclass correlation coefficient; IQR, interquartile range; κ, Cohen's kappa. Test–retest repeatability of Safety was evaluated using raw percentage agreement and Cohen's κ with 95% confidence intervals. For non-Safety metrics, ICC refers to ICC(2,1), a two-way random-effects, absolute-agreement, single-measurement intraclass correlation coefficient. ICC ≥ 0.75 was prespecified as acceptable test–retest repeatability. Cohen's dz was calculated as the mean paired difference (Round 2 − Round 1) divided by the standard deviation of paired differences; absolute values summarize the magnitude of round-to-round change. BH-significant drift indicates the number of non-Safety metrics with a BH-adjusted paired Wilcoxon *P* value < 0.05.

Discussion

This guideline-mapped evaluation found high safe-response rates across six generative artificial intelligence services, while potentially unsafe responses occurred in every service and clustered in medication management and post-procedure triage. These domains are clinically important after advanced colorectal endoscopic resection because inappropriate antithrombotic advice may

alter bleeding or thrombotic risk, whereas failure to recognize major bleeding, severe pain, systemic illness, or deterioration may delay urgent assessment [12, 13]. The absence of a significant overall Safety difference should not be interpreted as equivalence or proof of clinical safety.

Clinical accuracy and information-quality measures varied across tested services; the study was not designed to select a preferred system. Some archived responses from the tested Claude access condition self-identified as Claude Code, but those statements do not establish the exact interface or runtime environment. These results describe the observed Claude access condition and should not be generalized to the performance of the general-purpose Claude chat service or the Anthropic model family. This uncertainty limits direct comparisons with the other consumer-facing services.

The information-quality instruments captured different constructs. DISCERN and the study-adapted 24-item EQIP-based instrument assessed different aspects of patient-information quality and usability, whereas the JAMA benchmark focused on authorship, attribution, disclosure, and currency [19–22]. EQIP was developed for written healthcare information; its adaptation to conversational AI responses and the absence of a complete item-level rubric and scoring sheets limit item-level auditability and comparability with studies using the original instrument. Low JAMA scores primarily indicate limited transparency and do not necessarily imply factual inaccuracy. GQS provided a broader judgment of overall usefulness and flow. Readability remained difficult across services, with grade-level indices exceeding typical patient-education targets; in procedural care, linguistic complexity may compound risk when patients must understand medication instructions, distinguish routine recovery from complications, or recognize symptoms requiring urgent assessment [14–17].

After question-cluster-aware inference and Holm correction, longer pooled responses were associated with higher empathy, DISCERN, and JAMA scores; no other response-length associations remained significant. These pooled analyses combine six service responses to the same 70 questions and therefore do not isolate within-service relationships or establish causality. The second observed assessment qualified the Round 1 comparisons: using the same locked question wording and reported access procedures did not produce consistently high question-level score agreement. Low ICC(2,1) estimates describe observed between-round score stability, not a controlled assessment of stochastic variation under independently verified unchanged backend models. Several outcomes also showed directional changes between rounds; notably, EQIP medians increased across all six service conditions (Supporting Information S7). These shifts cannot be attributed uniquely to response generation, scoring implementation, or platform changes.

Strengths included the 70-question clinical benchmark, locked evidence mapping, five-rater evaluation, paired service comparison, unsafe-response characterization, and a complete second observed assessment. Limitations include English-language investigator-generated questions rather than a frequency-weighted patient sample; lack of patient comprehension, action, or clinical outcome measures; sparse potentially unsafe events; limited generalizability of the tested Claude access condition; and incomplete independent verification of vendor-side models and configurations. Although the aggregate scoring dataset contained no missing rater scores, the complete 24-item study-adapted EQIP-based rubric, item-level scoring sheets, and detailed applicability rules were not recoverable and must be supplied or clarified for reproducibility. The available records document 10 practice questions and independent rating, but do not provide a formal calibration statistic or complete details of post-rating adjudication. Extreme Round 2

readability values also warrant verification against the archived response text and preprocessing procedure before the readability findings are interpreted.

Conclusions

Potentially important unsafe advice persisted despite high safe-response rates under the observed service configurations, and question-level scores were not consistently stable across the two observed rounds. Clinically supervised use and configuration-specific, repeated evaluation may be particularly important for antithrombotic medication and urgent post-procedural questions.

List of abbreviations

AI: Artificial intelligence

ARI: Automated Readability Index

BH: Benjamini–Hochberg

CHART: Chatbot Assessment Reporting Tool

CI: Confidence interval

CL: Coleman–Liau Index

DISCERN: Instrument for assessing the quality of consumer health information about treatment choices

EMR: Endoscopic mucosal resection

EQIP: Ensuring Quality Information for Patients

ESD: Endoscopic submucosal dissection

ESGE: European Society of Gastrointestinal Endoscopy

FKGL: Flesch–Kincaid Grade Level

FRES: Flesch Reading Ease Score

GFI: Gunning Fog Index

GQS: Global Quality Score

ICC: Intraclass correlation coefficient

IQR: Interquartile range

JAMA: Journal of the American Medical Association benchmark criteria

LLM: Large language model

LST: Laterally spreading tumor

SD: Standard deviation

SMOG: Simple Measure of Gobbledygook

WEO: World Endoscopy Organization

Declarations

Ethics approval and consent to participate

This study evaluated investigator-formulated questions based on publicly available health-information themes and AI-generated text. No human participants were recruited or contacted, no clinical intervention was performed, and no human tissue, clinical records, or identifiable individual-level health data were collected or analyzed. Patient-facing themes were paraphrased rather than copied verbatim. Accordingly, ethics approval was not required for this study.

Consent to Participate declaration: Not applicable.

Consent for publication： Not applicable.

Availability of data and materials

Supporting Information S1–S7 contains the CHART checklist, question-development and screening materials, final question set, reference answers and evidence mapping, additional statistical analyses including question-cluster-aware correlation results, verbatim Round 1

responses, and the detailed two-round stability analysis. The complete Round 2 response archive for the 70-question, six-service benchmark is provided as Additional file 1. The question-linked scoring data for both rounds, including metric-level summary scores, individual rater scores for clinician-rated outcomes, a variable dictionary, and derived Round 1 character-count and word-count measures, are provided as Additional file 2. Analysis scripts are available from the corresponding author on reasonable request. Exact backend model identifiers and some interface or session settings were not recorded contemporaneously and could not be reconstructed.

Competing interests

The authors declare that they have no competing interests.

Funding

This study was conducted as a self-funded project registered under the Foshan Self-funded Science and Technology Innovation Project of the Foshan Municipal Bureau of Science and Technology (project no. 2320001006765). No external financial support was received.

Authors' contributions

Author initials: ZW, Zhihong Wang; JL, Jin Lyu; HHe, Huixin He; HHu, Huiwen Huang; QL, Qinghua Lu; and YO, Yongwen Ouyang.

ZW: Conceptualization (equal); Methodology (equal); Investigation (lead), including construction of the initial question pool and development of the reference answers and guideline evidence mapping; Data curation (lead); Formal analysis (lead), including the primary statistical analysis and second-round stability analysis; Visualization (equal); Writing – original draft (lead); Writing – review and editing (equal).

JL: Conceptualization (equal); Methodology (equal); Investigation (equal), including construction and screening of the question pool and development of the reference answers and guideline evidence mapping; Data curation (equal); Validation (lead), including verification of the evidence mapping, scoring dataset, statistical analysis, and between-round stability analysis; Writing – original draft (supporting); Writing – review and editing (equal).

HHe: Investigation (equal), including independent screening of candidate questions and verification of the reference answers and guideline evidence mapping; Data curation (supporting); Validation (equal); Writing – review and editing (equal).

HHu: Investigation (equal), including independent screening of candidate questions and standardized querying of the six tested service conditions in both assessment rounds; Data curation (equal), including organization and quality control of the response records; Validation (equal); Writing – review and editing (equal).

QL: Investigation (equal), including verification of the reference answers and guideline evidence mapping and verification of the standardized querying procedures; Data curation (equal); Validation (equal), including verification of service information, prompting consistency, timestamps where retained, response completeness, and second-round data; Visualization (equal); Writing – review and editing (equal).

YO: Conceptualization (equal); Methodology (supporting); Supervision (lead), including expert review and adjudication of disagreements arising during question screening and guideline evidence mapping; Project administration (lead); Writing – review and editing (lead); Funding acquisition (lead).

All authors reviewed and approved the final version of the manuscript and agreed to be accountable for all aspects of the work.

Acknowledgements

We thank Yun Huang and Xiangling Jiang for their contribution to question-set review and classification; both consented to be acknowledged.

References

1. Chen SF, Alyakin A, Seas A, Yang E, Choi JJ, Lee JV, et al. LLM-assisted systematic review of large language models in clinical medicine. Nat Med. 2026;32(3):1152-1159. doi:10.1038/s41591-026-04229-5.

2. 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. doi:10.1001/jamanetworkopen.2024.57879.

3. CHART Collaborative. Reporting guideline for chatbot health advice studies: the CHART statement. JAMA Netw Open. 2025;8(8):e2530220. doi:10.1001/jamanetworkopen.2025.30220.

4. CHART Collaborative. Reporting guidelines for chatbot health advice studies: explanation and elaboration for the Chatbot Assessment Reporting Tool (CHART). BMJ. 2025;390:e083305. doi:10.1136/bmj-2024-083305.

5. Gallifant J, Afshar M, Ameen S, Aphinyanaphongs Y, Chen S, Cacciamani G, et al. The TRIPOD-LLM reporting guideline for studies using large language models. Nat Med. 2025;31(1):60-69. doi:10.1038/s41591-024-03425-5.

6. He Z, Yang L, Liang Z, Li X, Du J. Clinical outcomes and reporting quality of large language model interventions in practice: a systematic evidence map. NPJ Digit Med. 2026;9(1):672. doi:10.1038/s41746-026-02837-6.

7. Djinbachian R, Rex DK, Chiu HM, Fukami N, Aihara H, Bastiaansen BAJ, et al. International consensus on the management of large (≥20 mm) colorectal laterally spreading tumors: World Endoscopy Organization Delphi study. Dig Endosc. 2024;36(11):1253-1268. doi:10.1111/den.14826.

8. Ferlitsch M, Hassan C, Bisschops R, Bhandari P, Dinis-Ribeiro M, Risio M, et al. Colorectal polypectomy and endoscopic mucosal resection: European Society of Gastrointestinal Endoscopy (ESGE) Guideline—Update 2024. Endoscopy. 2024;56(7):516-545. doi:10.1055/a-2304-3219.

9. Pimentel-Nunes P, Libânio D, Bastiaansen BAJ, Bhandari P, Bisschops R, Bourke MJ, et al. Endoscopic submucosal dissection for superficial gastrointestinal lesions: European Society of Gastrointestinal Endoscopy (ESGE) Guideline—Update 2022. Endoscopy. 2022;54(6):591-622. doi:10.1055/a-1811-7025.

10. Kaltenbach T, Anderson JC, Burke CA, Dominitz JA, Gupta S, Lieberman D, et al. Endoscopic removal of colorectal lesions: recommendations by the US Multi-Society Task Force on Colorectal Cancer. Gastroenterology. 2020;158(4):1095-1129. doi:10.1053/j.gastro.2019.12.018.

11. Hassan C, Antonelli G, Dumonceau JM, Regula J, Bretthauer M, Chaussade S, et al. Post-polypectomy colonoscopy surveillance: European Society of Gastrointestinal Endoscopy (ESGE) Guideline—Update 2020. Endoscopy. 2020;52(8):687-700. doi:10.1055/a-1185-3109.

12. Veitch AM, Radaelli F, Alikhan R, Dumonceau JM, Eaton D, Jerrome J, et al. Endoscopy in patients on antiplatelet or anticoagulant therapy: British Society of Gastroenterology (BSG) and European Society of Gastrointestinal Endoscopy (ESGE) guideline update. Endoscopy. 2021;53(9):947-969. doi:10.1055/a-1547-2282.

13. Abraham NS, Barkun AN, Sauer BG, Douketis J, Laine L, Noseworthy PA, et al. American College of Gastroenterology-Canadian Association of Gastroenterology Clinical Practice Guideline: management of anticoagulants and antiplatelets during acute gastrointestinal

bleeding and the periendoscopic period. Am J Gastroenterol. 2022;117(4):542-558. doi:10.14309/ajg.0000000000001627.

14. Siu AHY, Gibson DP, Chiu C, Kwok A, Irwin M, Christie A, et al. ChatGPT as a patient education tool in colorectal cancer—an in-depth assessment of efficacy, quality and readability. Colorectal Dis. 2025;27(1):e17267. doi:10.1111/codi.17267.

15. Yu J, Li M, Qin XZ, Gong L, Qin L, Teng Q, et al. Quality, transparency, and readability of AI chatbot responses for patient counseling in robot-assisted colorectal cancer resection: a cross-sectional evaluation. J Robot Surg. 2026;20(1):896. doi:10.1007/s11701-026-03874-9.

16. Maida M, Papaefthymiou A, Gupta S, Voiosu T, Lau LHS, Baraldo S, et al. Performance of large language models in addressing patient queries on colorectal cancer screening in different languages: an international study across 28 countries. Dig Liver Dis. 2026;58(2):250-257. doi:10.1016/j.dld.2025.11.026.

17. Ellison IE, Oslock WM, Abdullah A, Wood L, Thirumalai M, English N, et al. De novo generation of colorectal patient educational materials using large language models: prompt engineering key to improved readability. Surgery. 2025;180:109024. doi:10.1016/j.surg.2024.109024.

18. Peng W, Feng Y, Yao C, Zhang S, Zhuo H, Qiu T, et al. Evaluating AI in medicine: a comparative analysis of expert and ChatGPT responses to colorectal cancer questions. Sci Rep. 2024;14(1):2840. doi:10.1038/s41598-024-52853-3.

19. 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. doi:10.1136/jech.53.2.105.

552 20. Moult B, Franck LS, Brady H. Ensuring quality information for patients: development and
553 preliminary validation of a new instrument to improve the quality of written health care
554 information. Health Expect. 2004;7(2):165-175. doi:10.1111/j.1369-7625.2004.00273.x.
555 21. Silberg WM, Lundberg GD, Musacchio RA. Assessing, controlling, and assuring the quality
556 of medical information on the Internet: caveant lector et viewor—let the reader and viewer
557 beware. JAMA. 1997;277(15):1244-1245. doi:10.1001/jama.1997.03540390074039.
558 22. Bernard A, Langille M, Hughes S, Rose C, Leddin D, Veldhuyzen van Zanten S. A
559 systematic review of patient inflammatory bowel disease information resources on the World
560 Wide Web. Am J Gastroenterol. 2007;102(9):2070-2077. doi:10.1111/j.1572-
561 0241.2007.01325.x.
562 23. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for
563 reliability research. J Chiropr Med. 2016;15(2):155-163. doi:10.1016/j.jcm.2016.02.012.
564 24. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful
565 approach to multiple testing. J R Stat Soc Series B Stat Methodol. 1995;57(1):289-300.
566 doi:10.1111/j.2517-6161.1995.tb02031.x.
567 25. Flesch R. A new readability yardstick. J Appl Psychol. 1948;32(3):221-233.
568 doi:10.1037/h0057532.
569 26. Kincaid JP, Fishburne RP Jr, Rogers RL, Chissom BS. Derivation of new readability
570 formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for
571 Navy enlisted personnel. Research Branch Report 8-75. Millington (TN): Naval Technical
572 Training Command; 1975. Available from: https://apps.dtic.mil/sti/pdfs/ADA006655.pdf.
573 27. McLaughlin GH. SMOG grading—a new readability formula. J Read. 1969;12(8):639-646.

28. Smith EA, Senter RJ. Automated readability index. AMRL-TR-66-220. Wright-Patterson Air Force Base (OH): Aerospace Medical Research Laboratories; 1967. Available from: https://apps.dtic.mil/sti/pdfs/AD0667273.pdf.

29. Coleman M, Liau TL. A computer readability formula designed for machine scoring. J Appl Psychol. 1975;60(2):283-284. doi:10.1037/h0076540.

30. Gunning R. The technique of clear writing. New York (NY): McGraw-Hill; 1952.

Figure Legends

Figure 1. Study design and evaluation workflow. A 190-question candidate pool was screened to 70 locked, guideline-mapped questions across seven clinical domains and three risk tiers. Eighteen authoritative sources supported question-specific reference answers and safety anchors. Each question was submitted independently to six tested services during two observed rounds (420 responses per round). Five raters assessed clinical outcomes, and readability was evaluated programmatically. Round 1 comparisons used Cochran's Q, Friedman and paired tests as appropriate; repeatability was summarized by Safety agreement/Cohen's κ and non-Safety ICC(2,1). Actual backend identities could not be independently verified across rounds.

Figure 2. Distribution of Safety, Accuracy, and Empathy across the six generative artificial intelligence services. The Safety panel shows the number and percentage of responses classified as safe or as having a potential safety concern (the binary unsafe category used in the analysis) for each service (n = 70). Accuracy and Empathy panels show question-level values with boxplots summarizing the median and interquartile range; white diamonds indicate means.

Figure 3. Distribution of DISCERN, the adapted Ensuring Quality Information for Patients score, Journal of the American Medical Association benchmarks, and Global Quality Score across the six services. Points represent question-level values; boxplots summarize the median and interquartile range, and white diamonds indicate means. Higher values indicate better performance for all four measures.

Figure 4. Distribution of six readability indices across the six services. Points represent question-level values; boxplots summarize the median and interquartile range, and white diamonds indicate means. Higher Flesch Reading Ease indicates easier reading; higher Automated Readability Index, Coleman-Liau Index, Flesch-Kincaid Grade Level, Gunning Fog

Index, and SMOG values indicate greater reading difficulty. Red dashed lines show the reference thresholds used for visual comparison.

Figure 5. Correlation matrix of response length and evaluation metrics. Response length was quantified as character count for 420 Round 1 responses. Correlations involving binary Safety (0 = safe, 1 = potentially unsafe) used point-biserial Pearson correlation; all other pairs used Spearman rank correlation. Ninety-five percent CIs were estimated using 3,000 question-level bootstrap resamples preserving the six-service structure; P values were obtained from cluster-robust Wald tests with question ID as the clustering unit and Holm-adjusted across the 91 unique pairwise correlations displayed. Red indicates positive and blue indicates negative correlations. *Adjusted $P < 0.05$, **adjusted $P < 0.01$, and ***adjusted $P < 0.001$.

Supplementary information and additional files

Supporting Information S1. CHART checklist.

Supporting Information S2. Question-development protocol and screening log for 190 candidate questions.

Supporting Information S3. Final 70 English patient-facing questions with clinical-domain and risk classification.

Supporting Information S4. Reference answers, minimum essential content, unsafe-response anchors, and evidence mapping.

Supporting Information S5. Additional statistical analyses, including inter-rater agreement, pairwise service comparisons, unsafe-response distribution by clinical domain, and question-cluster-aware correlation analyses.

Supporting Information S6. Verbatim Round 1 question-and-answer records for all six tested service conditions.

Supporting Information S7. Detailed empirical between-round stability analysis, including Safety agreement, intraclass correlation coefficients, paired round comparisons, and readability stability.

Additional file 1. Complete structured archive of the 420 Round 2 responses (70 questions × six tested service conditions).

Additional file 2. Two-round scoring dataset and data dictionary, including metric-level scores and individual rater scores for clinician-rated outcomes.

Supplementary Files

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

- Additionalfile1Round2ResponseArchiveCorrected.docx
- Additionalfile2TwoRoundScoringDataandDictionaryCorrected.xlsx
- SupportingInformationS1Corrected.docx
- SupportingInformationS2Corrected.docx
- SupportingInformationS3Corrected.docx
- SupportingInformationS4Corrected.docx
- SupportingInformationS5Corrected.docx
- SupportingInformationS6Corrected.docx
- SupportingInformationS7Corrected.docx