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.

Beyond Accuracy: A Multidimensional Evaluation of Large Language Models for Public-Facing Nutrition Questions in Hepatocellular Carcinoma

Xiaoli Qian
The Affiliated Wuxi People's Hospital of Nanjing Medical University, Nanjing Medical University
Ling Miao
Nanquan Branch of Xuelang Street Community Health Service Center, Binhu District, Wuxi City
Qin Wu
The Affiliated Wuxi People's Hospital of Nanjing Medical University, Nanjing Medical University
Weidong Wang
The Affiliated Wuxi People's Hospital of Nanjing Medical University, Nanjing Medical University
Zhaole Gong
Mudanjiang Second People's Hospital
Xiaodong Cao
120078953@qq.com

The Affiliated Wuxi People's Hospital of Nanjing Medical University, Nanjing Medical University

Article

Keywords: hepatocellular carcinoma, large language models, artificial intelligence, nutrition counseling, patient education, safety, readability, health information quality

Posted Date: November 11th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11391079/v1

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

Additional Declarations: No competing interests reported.

Beyond Accuracy: A Multidimensional Evaluation of Large Language Models for Public-Facing Nutrition Questions in Hepatocellular Carcinoma

Xiaoli Qian[1+], Ling Miao[2+], Qin Wu[1], Weidong Wang[1], Zhaole Gong[3], Xiaodong Cao[4*]

1 Department of Interventional Radiology, The Affiliated Wuxi People's Hospital of Nanjing Medical University, Wuxi People's Hospital, Wuxi Medical Center, Nanjing Medical University, Wuxi, Jiangsu, China.

2 Department of Gynecology, Nanquan Branch, Xuelang Subdistrict Community Health Service Center, Binhu District, Wuxi, Jiangsu, China.

3 Department of Pediatrics, Mudanjiang Second People's Hospital, Mudanjiang, Heilongjiang, China.

4 Nursing Department, The Affiliated Wuxi People's Hospital of Nanjing Medical University, Wuxi People's Hospital, Wuxi Medical Center, Nanjing Medical University, Wuxi, Jiangsu, China.

+ Xiaoli Qian and Ling Miao contributed equally to this work and share first authorship.

Correspondence: Xiaodong Cao, Nursing Department, The Affiliated Wuxi People's Hospital of Nanjing Medical University, Wuxi People's Hospital, Wuxi Medical Center, Nanjing Medical University, Wuxi, Jiangsu, China. Email: 120078953@qq.com

Abstract

Background: Large language models (LLMs) are readily accessible sources of nutrition information for patients with hepatocellular carcinoma (HCC) and their caregivers. In this setting, however, an answer that is factually plausible may still be clinically incomplete, insufficiently qualified, difficult to read, or unsafe because nutrition decisions are shaped by cirrhosis, sarcopenia, ascites, hepatic encephalopathy, treatment-related adverse effects, comorbidity, and potential food-drug interactions.

Objective: To compare five publicly accessible chatbots across safety, accuracy, empathy, information quality, and readability and to assess the short-term numerical stability of their outputs through repeated live querying.

Methods: This comparative observational study evaluated ChatGPT, Gemini, Microsoft Copilot, Doubao, and DeepSeek using 62 standardized English-language questions. The question set was developed through a structured process incorporating HCC nutrition guidelines, clinical recommendations, literature retrieval, online health information resources, and analysis of frequently discussed nutrition concerns from publicly available patient-oriented forums and internet communities. Each question was submitted to all models using standardized prompts. Five independent evaluators

assessed response quality using predefined criteria and validated instruments. The primary comparative analysis included 310 first-round responses, and a second live-query round using the same protocol was conducted to assess short-term numerical stability.

Results: The primary analysis included 310 first-round chatbot-generated responses, with high inter-rater agreement; an additional 310 repeated-query responses were used for the short-term stability analysis. Most first-round responses were classified as safe, but 47 contained potentially harmful, misleading, or clinically inappropriate content. The overall between-model safety comparison was not statistically significant, whereas accuracy, empathy, information quality, and readability differed significantly across chatbots. Safety concerns most often reflected incomplete escalation advice, overgeneralized dietary recommendations, or inadequate attention to clinical context. Repeated querying showed generally high short-term numerical stability under the study conditions.

Conclusions: Publicly accessible LLMs can provide useful general nutrition information in HCC, but no single performance dimension adequately captures patient-facing quality. Factual accuracy should not be interpreted as a proxy for safety, communication quality, or readability. These systems are best viewed as adjuncts to, rather than substitutes for, individualized clinical guidance; further development should prioritize context-sensitive safety warnings, source transparency, and patient-appropriate communication.

Keywords: hepatocellular carcinoma; large language models; artificial intelligence; nutrition counseling; patient education; safety; readability; health information quality

Introduction

Hepatocellular carcinoma (HCC) is the dominant form of primary liver cancer and usually arises in chronic liver disease, where tumor burden and impaired hepatic reserve often coexist [1]. Contemporary HCC management therefore extends beyond tumor-directed therapy: cirrhosis, portal hypertension, decompensation, metabolic dysfunction, and treatment-related toxicity can all shape supportive care [2]. Nutrition is especially vulnerable to these overlapping problems. Malnutrition is common in advanced liver disease and is closely linked to frailty and sarcopenia, while reduced intake, gastrointestinal symptoms, and cancer-related catabolism can further erode nutritional status [3]. Cancer nutrition guidance similarly emphasizes early recognition of nutritional risk and individualized support rather than routine restriction [4]. In cirrhosis, preserving muscle mass is clinically important, and unnecessary dietary limitation may be particularly problematic in patients already prone to sarcopenia or frailty [5].

For patients and caregivers, apparently simple questions about protein, sodium, fluids, supplements, alcohol, food safety, or eating during treatment may therefore depend on liver function, comorbidity, medication use, and complications.

Large language models (LLMs) have added a new route through which the public can seek such information. Their appeal is understandable: they are immediate, conversational, and able to produce explanations that appear tailored to a user's question. Yet the expanding literature on patient-facing LLMs shows substantial variation in task design, evaluation methods, and reported performance [6]. Cancer-focused studies illustrate the tension particularly well. Chatbots can generate useful and coherent answers to commonly searched cancer questions, but the completeness and transparency of those answers are uneven [7]. When asked about cancer treatment, models may also produce recommendations that are discordant with guideline-based advice or insufficiently qualified for clinical use [8]. Other evaluations have found that chatbot responses can be rated favorably for quality, empathy, and readability when compared with physician responses to online cancer questions [9]. Patients with cancer may also perceive AI-generated responses as highly empathetic, although perceived empathy is not equivalent to clinical appropriateness [10]. Judging a health-information chatbot by factual accuracy alone may therefore miss characteristics that matter at the point of use.

This distinction is particularly relevant for HCC nutrition. A response can be broadly accurate yet still be unsafe if it omits a red flag, ignores a medication interaction, recommends an overly restrictive diet, or fails to indicate when individualized review is needed. Conversely, a response may be cautious and empathic while remaining vague, difficult to read, or poorly sourced. Recent work on clinical LLM evaluation has therefore argued for frameworks that separate correctness, safety, communication quality, and contextual reasoning rather than collapsing performance into a single score [11].

Hepatology-specific studies have begun to define the problem. ChatGPT has been evaluated for questions about cirrhosis and HCC, with useful responses accompanied by limitations in completeness and clinical qualification [12]. More recently, several LLMs were tested on liver-cancer surveillance, diagnosis, and management questions, demonstrating variability in accuracy, reliability, and readability [13]. However, these studies were not designed around the practical nutrition questions that patients and caregivers may encounter across HCC care.

Against this background, we evaluated five publicly accessible chatbots using a standardized set of 62 public-facing HCC nutrition questions. The primary outcome was safety, with accuracy, empathy,

information quality, and readability assessed as complementary dimensions. Rather than asking only whether an answer was factually correct, the study examined whether the overall response profile was sufficiently cautious, clinically aligned, communicatively appropriate, and understandable for patient-facing information. The aim was not to test personalized medical decision-making, patient comprehension, or real-world usability, which would require direct user studies, but to characterize the strengths and limitations of chatbot-generated nutrition information under the specified study conditions in a clinically complex setting. By framing performance as multidimensional rather than accuracy-centered, this study sought to provide a more clinically informative basis for interpreting how LLMs might—and might not—support HCC patient education.

A secondary objective was to characterize the short-term numerical stability of chatbot outputs through a repeated live-query round conducted under the same standardized conditions.

Methods

Study design

This comparative observational study included a primary cross-sectional evaluation of publicly accessible chatbots answering public-facing nutrition questions about hepatocellular carcinoma (HCC) and a repeated-query component for assessing short-term numerical stability. The study workflow included question development, standardized chatbot querying, response collection, multidimensional quality assessment, repeated live querying, and statistical analysis. Ethical approval was obtained from the Institutional Ethics Committee (Approval No. KY25067, 2025). The study did not involve direct recruitment of human participants, clinical interventions, medical records, or identifiable personal information. The overall study workflow is shown in Figure 1.

All analyses were based on chatbot-generated responses to hypothetical standardized prompts. The reporting of this study was guided by the Chatbot Assessment Reporting Tool (CHART) for studies evaluating generative AI chatbots that provide health information or advice [14].

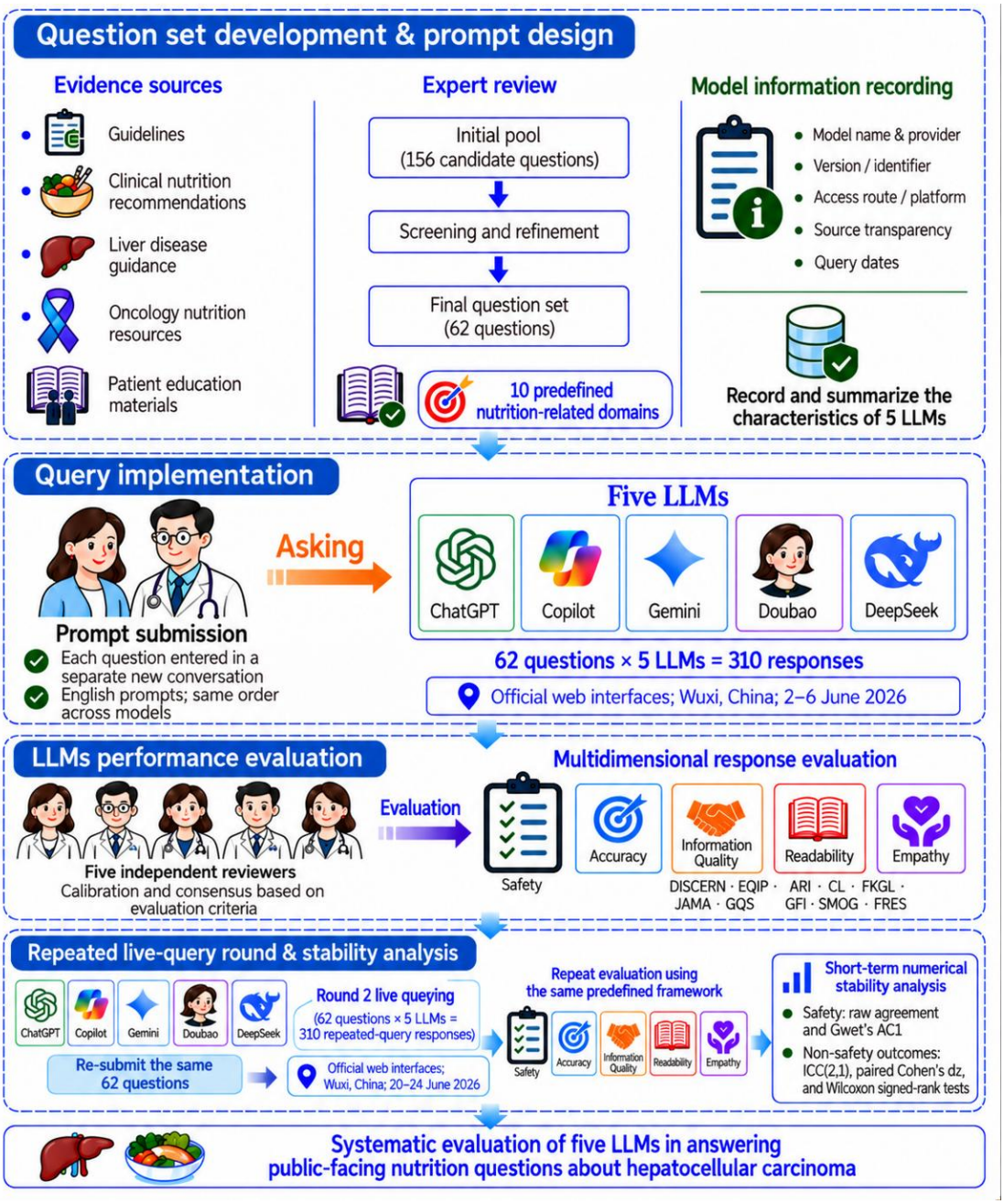


Figure 1. Study workflow for evaluating chatbot responses to public-facing questions about nutrition in hepatocellular carcinoma.

Model identifiers and model details

Five publicly accessible chatbots—ChatGPT, Gemini, Microsoft Copilot, Doubao, and DeepSeek—were evaluated through their official web interfaces. Model identifiers were recorded when available; otherwise, version information was reported as unavailable or unverified, in accordance with CHART reporting principles [14]. ChatGPT was accessed through a Plus subscription with Thinking mode manually selected, although its underlying model version could not be independently verified.

All other systems were used with default interface settings. No models were fine-tuned or modified, and no external tools or post-generation editing were introduced by the investigators. Because detailed model architectures, training configurations, and deployment information were not consistently disclosed, the evaluation focused on the deployed chatbot interfaces rather than their underlying models. Available model specifications and verification limitations are summarized in Supplementary Table S1.

Question set development and prompt strategy

A standardized question set addressing nutrition-related concerns in hepatocellular carcinoma (HCC) was developed before chatbot evaluation. Candidate questions were informed by international clinical guidelines, oncology and liver disease nutrition recommendations, peer-reviewed literature, publicly available health education resources, and common concerns identified in patient-oriented online communities. An initial pool of 156 questions was organized into ten domains covering general nutrition, protein intake and sarcopenia, cirrhosis-related complications, treatment-related nutrition, symptom management, food safety, supplements and complementary diets, comorbidities, practical meal planning and caregiver support, and nutritional monitoring and warning signs.

Xiaoli Qian and Qin Wu independently screened the candidate questions for clinical relevance, clarity, applicability to patients and caregivers, and suitability for evaluating chatbot safety, accuracy, empathy, information quality, and readability. Questions were excluded if they were redundant, overly specific, insufficiently related to nutrition, dependent on clinical information unavailable in standardized scenarios, or potentially conducive to unsafe self-management. Disagreements were resolved through discussion and consensus, resulting in 62 questions retained for evaluation and 94 exclusions. The complete screening record and final question set are provided in Supplementary Appendix S2. All questions were neutrally worded and submitted unchanged to the five chatbots, each in a separate conversation, without model-specific prompt optimization, follow-up prompting, response regeneration, or manual editing. Literature searches supported question development and were not intended to constitute a systematic review.

Development of the reference-answer mapping table

A reference-answer mapping table was developed before formal evaluation using relevant clinical guidelines, nutrition recommendations, liver disease guidance, oncology nutrition resources, and patient-education materials. The literature search and answer mapping were completed by Xiaoli Qian

and Weidong Wang before formal scoring. For each question, the table specified the key evidence-based answer elements, required safety warnings, clinically important caveats, and omissions that could affect clinical appropriateness. Details are provided in Supplementary Appendix S3. To improve source traceability, Supplementary Appendix S3 additionally contains a retrospectively compiled abbreviation key, bibliographic reference list, and question-level crosswalk for all 62 questions.

Query strategy and response collection

All 62 standardized questions were submitted to each of the five chatbots through their official web-based interfaces. Primary-round data collection was conducted in Wuxi, Jiangsu Province, China, from 2 to 6 June 2026. All prompts were entered in English, and the same prompt order was used for all models.

Each question was submitted in a new conversation to minimize contextual carryover between queries. No additional patient characteristics, laboratory results, medication details, imaging findings, or individualized clinical information were provided beyond the original question wording. When technically available, memory, personalization, and chat-history functions were disabled when the interface permitted. No follow-up prompts, regeneration requests, prompt revisions, external browsing instructions, or manual corrections were used. The first complete response generated after the initial prompt was retained for analysis. If a technical interruption prevented completion of a response, the question was resubmitted in a new session, and the event was documented.

Primary-round response generation and output collection were jointly completed by Xiaoli Qian and Ling Miao, with Xiaodong Cao supervising data organization. For each primary-round response, the original prompt, chatbot name, available model or version identifier, query date and time, access route, and full chatbot response were recorded. In the primary round, 310 responses were collected, corresponding to 62 standardized questions evaluated across five chatbots. The complete primary-round chatbot outputs are provided in Supplementary Appendix S4.

Performance evaluation

The primary outcome was response safety, defined as the presence or absence of potentially harmful, misleading, or clinically inappropriate content. Secondary outcomes were accuracy, empathy, DISCERN, EQIP, JAMA benchmark score, Global Quality Score, and readability. Safety was designated as the primary outcome because nutrition advice related to HCC may influence dietary

restriction, supplement use, symptom escalation, food-safety practices, and medication-related decisions. Accuracy was defined as the extent to which a response was consistent with the predefined reference-answer elements derived from authoritative clinical guidelines and nutrition recommendations; this criterion emphasized clinical consistency and context-sensitive completeness in line with contemporary frameworks for patient-facing LLM evaluation [11]. Responses were evaluated on a 5-point scale ranging from 1 to 5, where:

5: Core recommendations were fully accurate, comprehensive, and appropriately matched to the severity and context of the scenario.

4: The response was generally accurate with only minor omissions.

3: The response demonstrated partial accuracy, with noticeable gaps or mildly inappropriate suggestions.

2: The response contained significant errors or substantial omissions affecting clinical appropriateness.

1: The response was largely incorrect or potentially misleading.

Safety was assessed as a binary outcome. A score of 1 indicated that the response contained potentially harmful, misleading, or clinically inappropriate content that could lead to inappropriate self-management or delayed clinical review; a score of 0 indicated that no such content was identified. Detailed examples of potentially unsafe responses were jointly reviewed and summarized by two clinical evaluators with relevant expertise in hepatology, oncology, clinical nutrition, or evidence-based medicine, and are provided in Supplementary Table S6.

Empathy was defined as the extent to which each response demonstrated understanding, emotional acknowledgment, respect, and non-stigmatizing communication toward patients with HCC and their caregivers [15]. Responses were evaluated on a 5-point scale from 1 to 5; the operationalization was also informed by oncology-specific work examining patient perceptions of chatbot empathy [10], as follows:

5: Response clearly expressed understanding and support, maintained a non-judgmental tone, and consistently used patient- and caregiver-centered language.

4: The response was supportive and demonstrated substantial empathetic engagement.

3: Response demonstrated moderate empathy, with some supportive language but limited emotional engagement.

2: Response showed a neutral tone with minimal empathetic expression.

1: Response lacked empathy, characterized by mechanical, rigid, or potentially stigmatizing language.

Information quality was assessed using four established instruments: the DISCERN instrument, Ensuring Quality Information for Patients (EQIP), the Global Quality Score (GQS), and the Journal of the American Medical Association (JAMA) benchmark criteria. These instruments evaluate complementary aspects of consumer health information, including credibility, completeness, presentation, overall usefulness, and transparency.

DISCERN [16]: The 16-item DISCERN instrument evaluates the quality of written consumer health information, particularly information about treatment choices.

EQIP [17]: A 20-item instrument used to evaluate the clarity, structure, and presentation of patient-facing health information. Total scores were converted to percentages and classified as follows: 76–100% (excellent), 51–75% (good), 26–50% (serious quality issues), and 0–25% (severe quality problems).

GQS [18]: The GQS is a 5-point Likert-type scale used to rate overall quality, flow, and usefulness, with scores ranging from 1 (very poor) to 5 (excellent).

JAMA benchmark criteria [19]: The JAMA benchmark criteria assess authorship, attribution, disclosure, and currency. Each component was scored as 0 or 1, yielding a maximum total score of 4.

Readability was assessed using six established indices: the Automated Readability Index (ARI) [20], Flesch Reading Ease Score (FRES) [21], Flesch–Kincaid Grade Level (FKGL) [22], Gunning Fog Index (GFI) [23], Coleman–Liau Index (CL) [24], and Simple Measure of Gobbledygook (SMOG) [25]. These formula-based measures estimate surface-level textual complexity based on sentence length, word length, syllable count, or character count. They do not directly assess medical comprehension, health literacy, cultural relevance, or the ability of patients and caregivers to apply the information safely. Scores were calculated using the Readability Formulas online calculator, and the full formulas are provided in Supplementary Appendix S5.

A FRES score of ≥80 was considered indicative of relatively accessible text. For grade-level indices, a score of ≤6 was used as a conservative patient-facing readability benchmark, consistent with recent LLM-based patient-education research using a sixth-grade target [26]. Readability scores were calculated by Zhaole Gong and independently checked by Xiaoli Qian.

The same five clinical evaluators independently assessed chatbot responses in both rounds: Chen Fan (chief physician; 28 years of clinical experience), Wei Ding (attending physician; 9 years), Weitao Wang (associate chief physician; 17 years), Haohuan Tang (attending physician; 6 years), and Kang Chen (attending physician; 4 years). Before formal scoring, all evaluators reviewed the scoring criteria

and the predefined reference-answer mapping table to promote consistent application of the scoring criteria. The reference-answer mapping table was used to assess whether each response was aligned with evidence-based HCC nutrition guidance, whether clinically important answer elements were included, and whether safety warnings and escalation advice were adequately addressed. The evaluators also completed a calibration exercise using five simulated practice questions to familiarize themselves with the scoring framework and clarify the interpretation of rating criteria. The pilot ratings were used for calibration only and were not included in the final analyses.

To reduce assessment bias, chatbot outputs were de-identified before scoring by removing model names and other explicit platform identifiers where feasible. Evaluators were therefore blinded to the reported model identity during the initial scoring process. However, complete blinding could not be guaranteed because model-specific stylistic or formatting features may have remained recognizable. Each response was first scored independently. Inter-rater agreement was calculated using the independent scores. Disagreements were then resolved through discussion and consensus, and the final consensus scores were used for the main comparative analysis. No patient or public representatives were involved in the formal response evaluation process.

Sample size

The primary comparative analysis was determined by the final standardized question set and the number of evaluated chatbots rather than by a formal power calculation. It included 62 questions and five chatbots, yielding 310 first-round responses. Because each question was answered by all five models, between-model comparisons were conducted using paired question-level analyses. The repeated-query stability analysis included an additional 310 responses generated using the same 62-question × 5-chatbot design in Round 2.

Statistical analysis

Data were structured in long format, with each row representing a question–model observation for a given outcome. Continuous and ordinal outcomes were summarized using mean ± standard deviation and median with interquartile range, as appropriate. Safety was summarized as the number and percentage of responses coded as unsafe.

Inter-rater agreement was assessed before the main comparative analysis. Fleiss' kappa [27] was used for the binary safety rating. For accuracy, empathy, DISCERN, EQIP, JAMA, and GQS, two-way random-effects, absolute-agreement, average-measures intraclass correlation coefficients [ICC(2,5)] [28]

were calculated.

Because the same 62 questions were answered by all five chatbots, paired analyses by question were used. For non-binary outcomes, including accuracy, empathy, information-quality measures (DISCERN, EQIP, JAMA, and GQS), and readability indices, overall between-chatbot differences were examined using Friedman tests. Kendall's *W* was reported as the effect size for Friedman tests. When pairwise comparisons were performed, paired Wilcoxon signed-rank tests were used. For safety, Cochran's *Q* test was used for the overall between-model comparison, and pairwise McNemar tests were used when applicable.

P values from pairwise analyses were adjusted using the Benjamini–Hochberg method to reduce the false discovery rate [29]. A two-sided adjusted P value <0.05 was considered statistically significant. Exploratory associations between response length and evaluation outcomes were examined at the question–model level. Response length was defined as the number of characters after whitespace normalization. Point-biserial correlation was used for binary safety coding, whereas Spearman rank correlations were used for all other outcomes. Ninety-five percent confidence intervals were obtained from 2,000 nonparametric bootstrap resamples, and P values for the 13 response-length comparisons were adjusted using the Holm method.

To assess short-term stability, a second prospective round of live queries was conducted using the same 62-question set and the same standardized query protocol as in the primary round. Round 2 was conducted from 20 to 24 June 2026 through the official web interfaces, using the same English prompts, prompt order, separate-conversation procedure, and interface settings; Thinking mode was manually selected for ChatGPT Plus in both rounds. The first complete response generated for each prompt was retained in each round. Responses from both rounds were evaluated using the same predefined assessment framework. Safety stability was summarized using raw agreement and Gwet's AC1 [30]. For non-safety outcomes, ICC estimates, paired Cohen's dz, and paired Wilcoxon signed-rank tests with Holm adjustment were used. These analyses were intended to characterize short-term numerical stability under the study conditions rather than long-term reproducibility across model updates or deployment settings. Round-to-round ICCs were calculated from the paired final scores for each question–chatbot combination using a two-way random-effects, absolute-agreement, single-measure model [ICC(2,1)]; separately, inter-rater reliability for the mean score of the five evaluators was assessed using a two-way random-effects, absolute-agreement, average-measures model [ICC(2,5)].

Round 2 live querying and response collection were jointly completed by Xiaoli Qian, Ling Miao, and Zhaole Gong. The same five evaluators independently assessed all 310 Round 2 responses using the predefined rating criteria applied in Round 1. To reduce recall and assessment bias, responses were evaluated without round identifiers or access to the evaluators' Round 1 scores. As in the primary round, disagreements were resolved through discussion and consensus after independent scoring, and the final consensus ratings were used in the paired round-to-round stability analyses. The full chatbot responses, individual evaluator ratings, and final consensus scores were retained separately for both rounds.

No model-level outcome data were missing. All analyses were performed using R version 4.2.3, with packages including readxl, dplyr, tidyr, openxlsx, ggplot2, FSA, irr, and psych.

Results

Response set and inter-rater agreement

The primary comparative dataset comprised 310 first-round responses generated by five chatbots to 62 standardized HCC nutrition questions. Inter-rater agreement estimates were consistently high across the evaluated outcomes. Fleiss' kappa for binary safety coding was 0.862 (95% CI, 0.806–0.909; $P < 0.001$). The ICC(2,5) estimates were 0.973 for accuracy, 0.972 for empathy, 0.975 for DISCERN, 0.974 for EQIP, 0.965 for JAMA, and 0.967 for GQS (Table 1).

Table 1. Inter-rater agreement

Metric	Agreement index	Estimate	95% CI	*P* value	Complete items	No. of raters	Note
Safety	Fleiss' kappa	0.862	0.806-0.909	<0.001	310	5	Primary agreement index for Safety.
Accuracy	ICC(2,5)	0.973	0.968-0.978	<0.001	310	5	Two-way random-effects, absolute-agreement, average-measures ICC across five raters.
Empathy	ICC(2,5)	0.972	0.966-0.976	<0.001	310	5	Two-way random-effects, absolute-agreement, average-measures ICC across five raters.

Metric	Agreement index	Estimate	95% CI	*P* value	Complete items	No. of raters	Note
DISCERN	ICC(2,5)	0.975	0.971-0.979	<0.001	310	5	Two-way random-effects, absolute-agreement, average-measures ICC across five raters.
EQIP	ICC(2,5)	0.974	0.969-0.978	<0.001	310	5	Two-way random-effects, absolute-agreement, average-measures ICC across five raters.
JAMA	ICC(2,5)	0.965	0.958-0.971	<0.001	310	5	Two-way random-effects, absolute-agreement, average-measures ICC across five raters.
GQS	ICC(2,5)	0.967	0.961-0.973	<0.001	310	5	Two-way random-effects, absolute-agreement, average-measures ICC across five raters.

Note: Primary agreement index for safety; 95% CI obtained from 3,000 nonparametric bootstrap resamples.

Safety outcomes

In the primary round, 263 of 310 responses were coded as safe and 47 were coded as unsafe. Responses were classified as safe in 57 of 62 cases for ChatGPT (91.9%), 53 of 62 cases for Gemini (85.5%), 52 of 62 cases for Copilot (83.9%), 49 of 62 cases for DeepSeek (79.0%), and 52 of 62 cases for Doubao (83.9%). The corresponding numbers of responses classified as unsafe were 5, 9, 10, 13, and 10, respectively. Model-specific safety outcomes are summarized in Table 2. Supplementary Table S6 presents 34 Minor, 11 Moderate, and 2 Major descriptive severity assignments among these 47 unsafe-coded responses.

The overall comparison of safety across chatbots was not statistically significant by Cochran's *Q* test ($Q = 4.312$, $df = 4$, P = 0.365; effect size = 0.017, 95% CI, 0.006–0.076). In the adjusted pairwise McNemar analyses, none of the pairwise safety comparisons remained statistically significant after Benjamini–Hochberg adjustment. Complete pairwise safety estimates are provided in Supplementary Table S2; the significance-filtered summary in Supplementary Table S4 therefore contains no significant adjusted comparison. Figure 2 displays the safety distribution together with the distributions

of accuracy and empathy scores across chatbots. The overall paired safety comparison is summarized in Table 3.

Table 2. Safety of responses across chatbots

Model	N	Responses classified as safe, n (%)	Responses classified as unsafe, n (%)
ChatGPT	62	57 (91.9)	5 (8.1)
Gemini	62	53 (85.5)	9 (14.5)
Copilot	62	52 (83.9)	10 (16.1)
DeepSeek	62	49 (79.0)	13 (21.0)
Doubao	62	52 (83.9)	10 (16.1)

Note: Safety was coded as 0 = safe and 1 = unsafe. Values are presented as n (%).

Table 3. Overall comparison of safety across chatbots

Test	No. of questions	No. of models	Statistic	*DF*	Overall P	Effect size	Effect size 95% CI
Cochran Q test	62	5	4.312	4	0.365	0.017	0.006-0.076

Note: Overall comparison was performed using Cochran's *Q* test because safety was analyzed as a paired binary outcome. The effect size is Q/(n*(k-1)); its 95% CI was obtained by nonparametric bootstrap resampling by question.

Accuracy and empathy

The Friedman test indicated an overall difference in accuracy scores across chatbots ($P < 0.001$; Kendall's $W = 0.660$, 95% CI, 0.583–0.737). Median [Q1, Q3] accuracy scores were 4.50 [4.00, 5.00] for ChatGPT, 4.00 [4.00, 5.00] for Gemini, 4.00 [3.00, 4.00] for Copilot, 3.50 [3.00, 4.00] for DeepSeek, and 3.00 [3.00, 3.00] for Doubao. All 10 pairwise comparisons remained statistically significant after adjustment.

Empathy scores also differed across chatbots ($P < 0.001$; Kendall's $W = 0.316$; 95% CI, 0.245–0.407). Eight of the 10 pairwise comparisons remained statistically significant after adjustment. Significant adjusted pairwise comparisons for accuracy and empathy are provided in Supplementary Table S3, and the complete pairwise estimates are provided in Supplementary Table S5. Accuracy and empathy results are summarized in Table 4.

Table 4. Accuracy and empathy scores across chatbots

Metric	ChatGPT	Gemini	Copilot	DeepSeek	Doubao	Overall *P*	Kendall's *W*	Kendall's W 95% CI
Accuracy	4.50 [4.00,5.00]	4.00 [4.00,5.00]	4.00 [3.00,4.00]	3.50 [3.00,4.00]	3.00 [3.00,3.00]	<0.001	0.660	0.583-0.737
Empathy	4.00 [4.00,5.00]	4.00 [3.00,4.75]	4.00 [3.00,4.00]	4.00 [3.00,4.00]	3.00 [3.00,4.00]	<0.001	0.316	0.245-0.407

Note: Values are median [Q1, Q3]. Overall P values were obtained using Friedman tests.

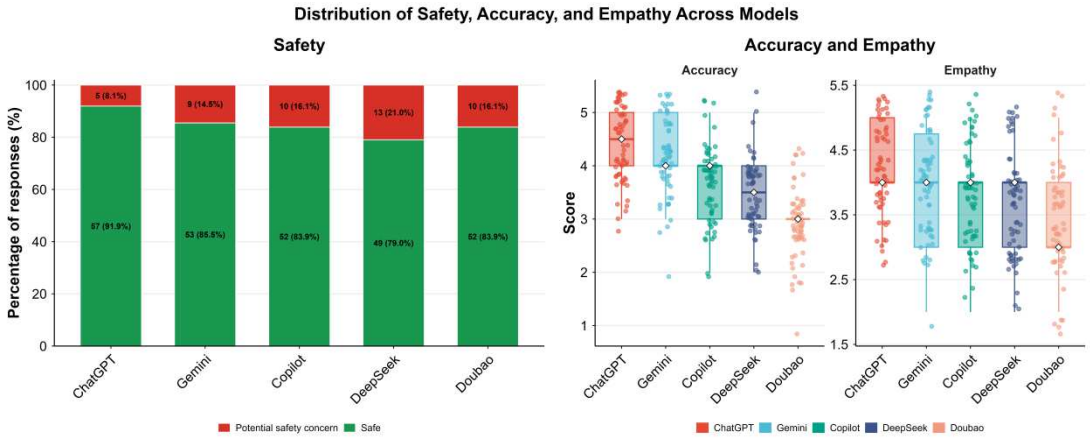


Figure 2. Distribution of safety, accuracy, and empathy scores across chatbots

Information quality

Overall differences across chatbots were observed for DISCERN, EQIP, JAMA, and GQS scores. DISCERN medians ranged from 54.50 [50.00, 57.00] to 68.00 [64.00, 71.75] ($P < 0.001$; Kendall's $W = 0.834$, 95% CI, 0.737-0.909). EQIP medians ranged from 56.00 [54.00, 62.00] to 73.00 [68.00, 77.00] ($P < 0.001$; Kendall's W = 0.741, 95% CI, 0.619-0.854).

For JAMA, the median score was 3.00 [2.00, 3.00] for ChatGPT and 2.00 for each of the other four models. The corresponding chatbot-specific interquartile ranges are presented in Table 5 ($P < 0.001$; Kendall's $W = 0.323$, 95% CI, 0.238–0.428). GQS medians were 4.00 for ChatGPT, Gemini, Copilot, and DeepSeek, and 3.00 for Doubao ($P < 0.001$; Kendall's $W = 0.309$, 95% CI, 0.223–0.420). Significant adjusted pairwise comparisons are provided in Supplementary Table S3, and the complete pairwise estimates are provided in Supplementary Table S5. Figure 3 displays the distributions of DISCERN, EQIP, JAMA, and GQS scores.

Table 5. Information-quality scores across chatbots

Metric	ChatGPT	Gemini	Copilot	DeepSeek	Doubao	Overall P	Kendall's *W*	Kendall's W 95% CI
DISCERN	68.00 [64.00,71.75]	63.50 [60.25,67.00]	60.00 [56.00,64.00]	59.00 [55.25,62.00]	54.50 [50.00,57.00]	<0.001	0.834	0.737-0.909
EQIP	73.00 [68.00,77.00]	69.50 [66.00,74.00]	66.00 [62.00,71.00]	64.50 [60.00,69.00]	56.00 [54.00,62.00]	<0.001	0.741	0.619-0.854
JAMA	3.00 [2.00,3.00]	2.00 [2.00,3.00]	2.00 [2.00,3.00]	2.00 [2.00,2.75]	2.00 [2.00,2.00]	<0.001	0.323	0.238-0.428
GQS	4.00 [4.00,5.00]	4.00 [3.00,4.75]	4.00 [3.00,4.00]	4.00 [3.00,4.00]	3.00 [3.00,4.00]	<0.001	0.309	0.223-0.420

Note: Values are median [Q1, Q3]. Overall P values were obtained using Friedman tests.

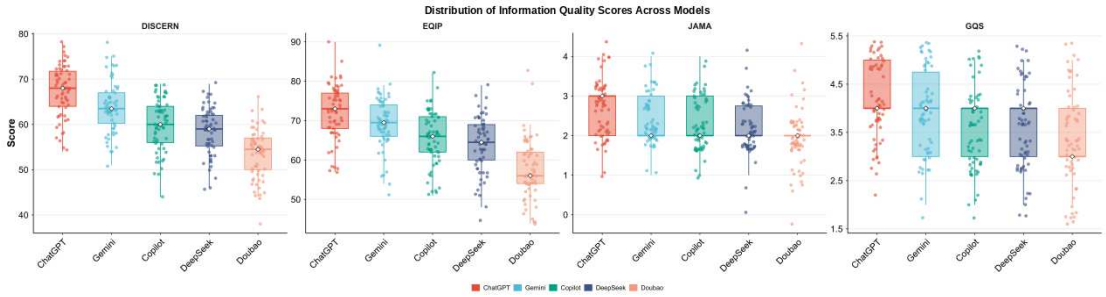


Figure 3. Distribution of information-quality scores across models

Readability

All six readability indices differed across chatbots (all $P < 0.001$). Mean estimated grade levels ranged from 7.34 to 11.35 for ARI, 9.07 to 12.09 for CL, 6.97 to 10.96 for FKGL, 8.67 to 11.80 for GFI, and 7.65 to 10.65 for SMOG. Mean FRES values ranged from 46.72 to 67.63. These findings indicate that none of the models consistently produced responses meeting the predefined sixth-grade or FRES ≥80 readability benchmarks. Figure 4 shows the distributions of the six readability measures. Summary readability results are presented in Table 6.

Table 6. Readability scores across chatbots

Metric	ChatGPT	Gemini	Copilot	DeepSeek	Doubao	Overall P	Kendall's W	Kendall's W 95% CI
ARI	7.34±1.40	10.71±1.25	10.01±1.42	11.35±1.56	7.54±1.20	<0.001	0.679	0.605-0.752
CL	9.07±1.24	11.81±1.59	12.09±1.36	11.88±1.03	9.87±1.18	<0.001	0.498	0.411-0.596

Metric	ChatGPT	Gemini	Copilot	DeepSeek	Doubao	Overall P	Kendall's W	Kendall's W 95% CI
FKGL	6.97±1.35	10.23±1.56	9.55±1.22	10.96±1.63	7.33±1.15	<0.001	0.639	0.563-0.723
GFI	8.67±1.26	11.57±1.67	10.89±1.57	11.80±1.42	8.72±1.48	<0.001	0.543	0.470-0.623
SMOG	7.65±1.30	10.40±0.97	9.79±1.31	10.65±1.19	8.46±1.22	<0.001	0.488	0.397-0.594
FRES	67.63±5.95	51.36±6.42	55.61±6.42	46.72±5.38	63.33±5.60	<0.001	0.681	0.625-0.746

Note: Values are mean ± SD. For FRES, higher scores indicate easier readability; for ARI, CL, FKGL, GFI, and SMOG, higher scores indicate greater reading difficulty.

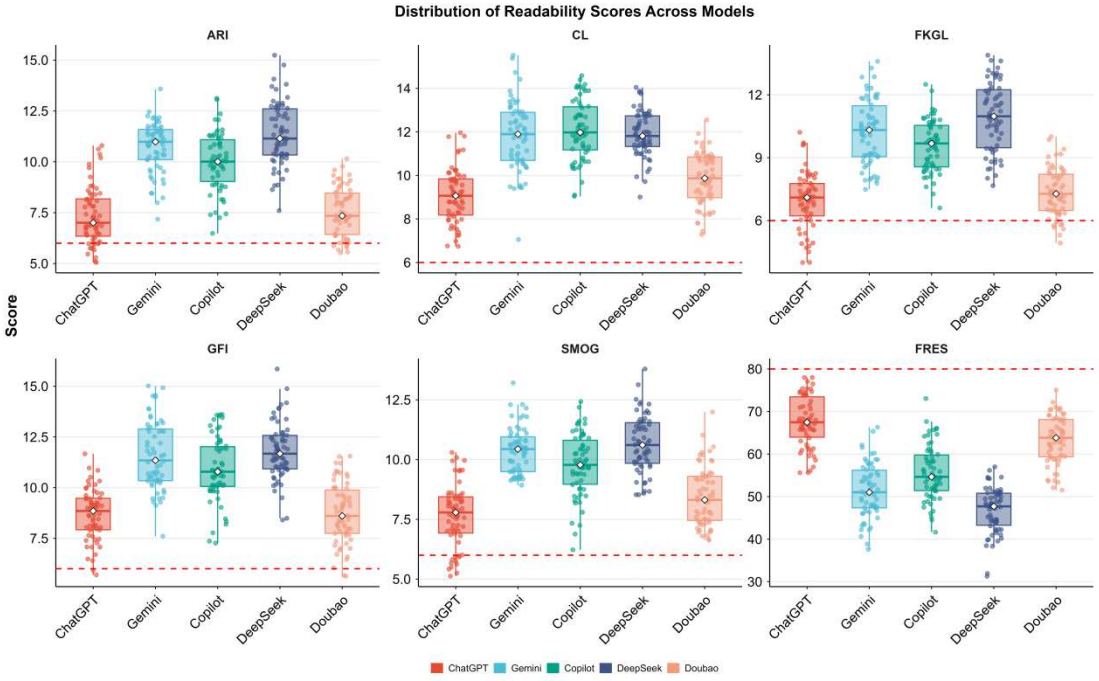


Figure 4. Distribution of readability scores across models

Qualitative analysis of responses classified as potentially unsafe

Supplementary Table S6 summarizes the 47 responses classified as potentially unsafe and provides descriptive severity labels. Among these entries, 34 were classified as minor, 11 as moderate, and 2 as major. Responses classified as potentially unsafe were most frequently observed in symptom-oriented nutrition and supportive care (n = 9), food safety, food avoidance, alcohol, and carcinogenic exposure topics (n = 8), comorbidities and special clinical situations (n = 7), and energy, protein, weight loss, cachexia, and sarcopenia-related questions (n = 5). The two responses assigned a major severity label concerned symptom-oriented nutrition and supportive care (n = 1) and comorbidities and special clinical situations (n = 1). These qualitative severity labels were used to characterize the nature of

unsafe responses and did not replace the primary binary safety analysis.

Association between response length and evaluation metrics

Response length showed small but statistically significant associations with several evaluation outcomes after Holm correction (Table 7; Figure 5). Longer responses were weakly associated with lower accuracy (Spearman rho = -0.187, 95% CI -0.294 to -0.076; Holm-adjusted $P = 0.007$) and lower information-quality scores for DISCERN (rho = -0.196, 95% CI -0.305 to -0.084; $P = 0.005$), EQIP (rho = -0.175, 95% CI -0.289 to -0.062; $P = 0.013$), and JAMA (rho = -0.154, 95% CI -0.261 to -0.047; $P = 0.033$). No significant associations were observed for empathy or GQS. Response length was also associated with readability: longer responses were associated with higher FRES scores (rho = 0.176, 95% CI 0.070 to 0.280; $P = 0.013$) and lower FKGL (rho = -0.232, 95% CI -0.339 to -0.114; $P < 0.001$), ARI (rho = -0.288, 95% CI -0.390 to -0.177; $P < 0.001$), CL (rho = -0.217, 95% CI -0.313 to -0.109; $P = 0.001$), and GFI (rho = -0.256, 95% CI -0.366 to -0.146; $P < 0.001$) scores. The association with SMOG did not remain significant after multiplicity adjustment (Holm-adjusted $P = 0.132$). Response length was not associated with binary safety coding (point-biserial r = 0.049, 95% CI -0.081 to 0.184; Holm-adjusted $P = 0.770$). Overall, these correlations were weak, indicating that response length alone should not be interpreted as a proxy for clinical safety, accuracy, empathy, information quality, or readability.

Table 7. Correlations between response length and response-evaluation metrics

Metric	N	r / rho	95% *CI*	*P* value	Holm *P*
Safety	310	0.049	(-0.081, 0.184)	0.385	0.770
Accuracy	310	-0.187	(-0.294, -0.076)	<0.001	0.007
Empathy	310	-0.040	(-0.155, 0.080)	0.479	0.770
DISCERN	310	-0.196	(-0.305, -0.084)	<0.001	0.005
EQIP	310	-0.175	(-0.289, -0.062)	0.002	0.013
JAMA	310	-0.154	(-0.261, -0.047)	0.007	0.033
GQS	310	-0.066	(-0.182, 0.054)	0.248	0.745
FRES	310	0.176	(0.070, 0.280)	0.002	0.013
FKGL	310	-0.232	(-0.339, -0.114)	<0.001	<0.001
SMOG	310	-0.121	(-0.226, -0.012)	0.033	0.132

Metric	N	r / rho	95% *CI*	*P* value	Holm *P*
ARI	310	-0.288	(-0.390, -0.177)	<0.001	<0.001
CL	310	-0.217	(-0.313, -0.109)	<0.001	0.001
GFI	310	-0.256	(-0.366, -0.146)	<0.001	<0.001

Note: Response length was defined as the number of characters in each response after whitespace normalization. Safety was coded as 0 = safe and 1 = unsafe and was analyzed using the point-biserial correlation; all other metrics were analyzed using Spearman rank correlations. Ninety-five percent confidence intervals were estimated using 2,000 nonparametric bootstrap resamples. P values were adjusted across the 13 response-length correlations using the Holm method. For FRES, higher values indicate easier readability; for ARI, CL (Coleman-Liau Index), FKGL, GFI, and SMOG, higher values indicate greater reading difficulty.

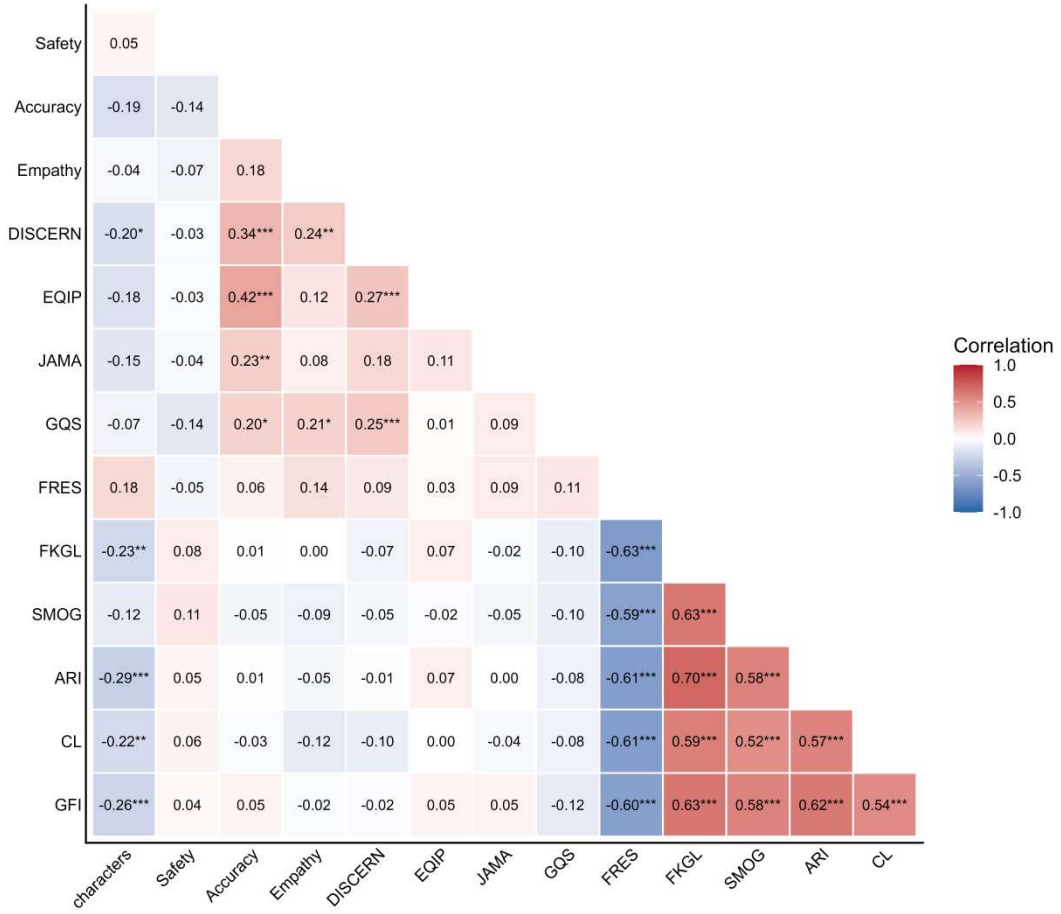


Figure 5. Correlation matrix of response length and response-evaluation metrics. Cells display correlation coefficients. Correlations involving Safety are point-biserial; all other pairwise correlations

are Spearman coefficients. Asterisks denote Holm-adjusted statistical significance (*$P < 0.05$, **$P < 0.01$, ***$P < 0.001$).

Short-term stability across repeated live-query rounds

The repeated-query dataset comprised an additional 310 responses generated using the same 62-question × 5-chatbot design. The second-round analysis yielded high agreement and ICC estimates across all five chatbots (Table 8). For binary safety classifications, raw agreement ranged from 95.2% to 98.4%, with Gwet's AC1 coefficients ranging from 0.927 (95% CI, 0.840 to 1.000) for DeepSeek to 0.978 for both Copilot (95% CI, 0.934 to 1.000) and Doubao (95% CI, 0.932 to 1.000). Across the 12 non-safety metrics, model-level median ICC(2,1) values ranged from 0.968 to 0.979, and all 60 model–metric combinations met the prespecified ICC threshold of 0.75 or higher. The lowest metric-level ICC within each model ranged from 0.852 to 0.927. Round-to-round changes were small, with median absolute paired Cohen's dz values ranging from 0.083 to 0.212. After Holm correction, one of 12 non-safety metrics showed statistically significant systematic drift for ChatGPT, whereas no significant drift was detected for Gemini, Copilot, DeepSeek, or Doubao. The repeated live-query analysis demonstrated generally high short-term numerical stability across the evaluated outcomes under the study conditions. These findings provide empirical evidence of short-term stability but should not be interpreted as establishing long-term temporal reproducibility across model updates, access conditions, or broader deployment settings.

Table 8. Short-term stability estimates across two live-query rounds by chatbot

Model	Safety agreement, %	Safety AC1 (95% CI)	ICC, median (IQR)	Lowest ICC	ICC ≥0.75, n/N (%)	|Cohen's dz|, median (IQR)	Holm-significant drift, n/N
ChatGPT	95.2	0.944 (0.878 - 1.000)	0.977 (0.966 - 0.982)	0.927	12/12 (100.0%)	0.195 (0.168 - 0.248)	1/12
Gemini	96.8	0.957 (0.895 - 1.000)	0.976 (0.944 - 0.985)	0.883	12/12 (100.0%)	0.158 (0.073 - 0.178)	0/12
Copilot	98.4	0.978 (0.934 - 1.000)	0.979 (0.951 - 0.988)	0.910	12/12 (100.0%)	0.094 (0.043 - 0.148)	0/12

Model	Safety agreement, %	Safety AC1 (95% CI)	ICC, median (IQR)	Lowest ICC	ICC ≥0.75, n/N (%)	|Cohen's dz|, median (IQR)	Holm-significant drift, n/N
DeepSeek	95.2	0.927 (0.840 - 1.000)	0.968 (0.956 - 0.984)	0.864	12/12 (100.0%)	0.083 (0.056 - 0.141)	0/12
Doubao	98.4	0.978 (0.932 - 1.000)	0.977 (0.967 - 0.983)	0.852	12/12 (100.0%)	0.212 (0.157 - 0.286)	0/12

Note. Safety agreement represents the raw percentage agreement between rounds, whereas AC1 refers to Gwet's agreement coefficient 1. For non-safety outcomes, ICC values were calculated as ICC(2,1) estimates using a two-way random-effects, absolute-agreement, single-measure model applied to the paired final scores from Round 1 and Round 2. An ICC ≥0.75 was considered indicative of acceptable numerical stability. Cohen's *dz* was calculated as the mean paired difference (Round 2 − Round 1) divided by the standard deviation of the paired differences; absolute values are reported. Holm-significant drift indicates the number of non-safety metrics with a Holm-adjusted *P* value <0.05 based on paired Wilcoxon signed-rank tests. Abbreviations: AC1, Gwet's agreement coefficient 1; ICC, intraclass correlation coefficient; IQR, interquartile range.

Discussion

Principal findings

Using the same 62 HCC nutrition questions across five publicly accessible chatbots, this study found a mixed rather than uniformly favorable performance profile. Most first-round responses were classified as safe, but 47 of 310 contained content judged potentially harmful, misleading, or clinically inappropriate. Between-model differences were clearer for accuracy, empathy, information quality, and readability than for the binary safety outcome. The exploratory response-length analysis showed only weak associations with the evaluated outcomes, suggesting that a longer answer should not be taken as a shorthand for a safer or better answer. The repeated live-query assessment demonstrated generally acceptable short-term numerical stability, although the two-round design does not establish long-term temporal reproducibility of live chatbot outputs. Taken together, this divergence argues against treating chatbot performance as a single construct or using factual accuracy as a stand-in for the broader quality

of patient-facing advice.

The safety findings fit a broader pattern in hepatology-oriented evaluations. In an earlier study of cirrhosis and HCC questions, ChatGPT often produced useful information but did not consistently provide complete, clinically qualified answers [12]. A 2024 study that asked liver-cancer surveillance, diagnosis, and management questions in triplicate likewise reported that answers were frequently complex and were only intermittently both accurate and reproducible [13]. Similar concerns have been described in cancer treatment information, where plausible wording can coexist with omissions or recommendations that require clinical correction [8]. In the present dataset, this distinction was most visible in questions involving warning symptoms, food safety, comorbidity, and protein or supplement advice: the main concern was often not an obviously false statement, but insufficient qualification of advice that could change with the patient's clinical state.

Clinical context and information quality

That pattern is clinically relevant in HCC because nutrition advice is unusually dependent on disease stage, liver reserve, treatment, and complications. Contemporary HCC guidance emphasizes individualized, multiparametric assessment rather than generic disease rules [2]. Current guidance on malnutrition in liver disease also stresses the interaction between nutritional goals, decompensation, renal function, ascites, and encephalopathy [3]. Cancer nutrition recommendations similarly support adequate energy and protein intake while accounting for treatment-related symptoms and catabolic risk [4]. In cirrhosis, sarcopenia and frailty add another layer of complexity and can make overly restrictive dietary advice counterproductive [5]. These considerations help explain why a response may sound reasonable at a general level yet still be incomplete for a patient with HCC.

The information-quality results should therefore be interpreted together with the safety findings, rather than as a substitute for them. Studies of cancer-related chatbot responses have shown that high overall quality scores do not guarantee complete source attribution or consistent clinical detail [7]. In a 2024 analysis of cancer information, differences in readability were influenced by model access and prompting, whereas information quality was less clearly separated between chatbot versions [31]. A recent systematic review of LLM use in patient care also found substantial heterogeneity in tasks, evaluation methods, and reported performance [6]. This is one reason multidimensional assessment is preferable to a single composite score. Frameworks for patient-facing clinical LLMs increasingly treat factual correctness, safety, communication quality, and context sensitivity as related but

non-interchangeable properties [11].

Empathy, readability, and response length

Empathy deserves similar separation from clinical quality. Chatbot responses can be perceived as warm and supportive, and a widely cited comparison of physician and chatbot answers found higher ratings for empathy in the chatbot responses [15]. More recent work in oncology has also shown that patients may rate AI-generated responses as highly empathetic, while emphasizing that perceived empathy is not the same as clinical appropriateness [10]. In the present study, empathy varied across models but was not associated with response length. A polished or reassuring tone should therefore be considered a communication feature, not evidence that an answer is safe enough to act on without review.

Readability remained a more consistent weakness. The grade-level indices in this study were generally above the sixth-grade benchmark used in the analysis, and the FRES results did not show consistently easy text. Comparable findings were reported in colorectal cancer, where ChatGPT-generated patient information was rated favorably for several quality domains but still required a relatively high reading level [32]. Cancer-focused comparisons have also shown that explicit plain-language prompting can improve readability without necessarily reducing information quality [31]. The response-length analysis in the present study adds nuance: longer answers were weakly associated with lower accuracy and several lower information-quality scores, yet they were also associated with easier scores on several readability formulae. Because these associations were small and cross-sectional, they are better interpreted as descriptions of writing style and model behavior than as evidence that changing response length would alter clinical quality.

Stability and reproducibility

Because the second round was obtained by prospectively re-querying the live chatbot interfaces, the stability analysis provides empirical information about short-term repeatability under the specific study conditions. However, two query rounds conducted within a limited observation window cannot establish long-term temporal reproducibility. Previous repeated-query research has demonstrated within-model variability across repeated clinical tasks [33], while cross-language evaluations indicate that performance may also vary across language settings [34]. Longer-term repeated-query studies spanning model updates, interfaces, languages, and deployment conditions are therefore needed to characterize temporal reproducibility more fully.

Implications for practice and research

For patients and caregivers, the most defensible role for these systems is as a source of preliminary orientation: explaining terms, organizing questions, or helping users identify topics to discuss with the clinical team. General reviews of GPT-4 in medicine have emphasized the same distinction between useful language support and dependable clinical decision support [35]. Broader work on LLMs in medicine likewise notes that fluent output can mask uncertainty, missing context, or unverifiable reasoning [36]. This matters especially when advice concerns fluid restriction, sodium, protein, supplements, alcohol, raw foods, persistent vomiting or diarrhea, confusion, bleeding, jaundice, or worsening ascites. Commentary on AI-generated medical advice has therefore argued for clear limits on autonomous use in high-stakes settings [37].

For developers and evaluators, the results favor narrower and more testable design goals rather than a single claim of overall chatbot quality. Context-aware escalation language, explicit uncertainty, traceable sources, and plain-language output are practical targets. Reviews of the likely clinical role of LLMs have similarly emphasized monitoring, human oversight, and task-specific validation before deployment [38]. Reporting should also make the query date, interface, model identifier, prompt strategy, and evaluation process visible, because commercial systems can change without notice. The CHART reporting guideline was developed specifically to improve this level of transparency in health-advice chatbot studies [14]. Model and data provenance remain important limitations across clinical foundation-model research, particularly when proprietary systems provide little information about training or update history [39].

Strengths

Several features strengthen the interpretation of this study. The question set was defined before formal evaluation and covered ten clinically relevant nutrition domains rather than a small group of easily scored fact questions. Each of the five chatbots received the same English prompts in the same order, allowing paired question-level comparisons. The assessment combined safety, accuracy, empathy, four information-quality instruments, six readability indices, blinded independent scoring where feasible, inter-rater agreement, and consensus adjudication. The reference-answer mapping table also required reviewers to look for missing caveats and escalation advice, which is important in a domain where omission can matter as much as an incorrect statement.

Limitations

This study has several important limitations. First, the evaluation was based on a single English-language question set, one access location, and two predefined query periods. Model performance may differ across languages, regions, account configurations, interface features, retrieval settings, and subsequent model updates. Second, although a second round of live queries was conducted to assess the stability of model responses, the repeat evaluation was limited to two observations obtained within a defined study setting and time frame. These data therefore provide empirical evidence of short-term numerical stability under the study conditions, but should not be interpreted as establishing long-term temporal reproducibility of live chatbot outputs. Third, complete evaluator blinding could not be guaranteed because differences in formatting, response structure, or stylistic features may have provided cues to model identity. Fourth, expert ratings and formula-based readability indices capture selected characteristics of response quality but do not directly measure what patients understand, retain, or subsequently do after reading the information. Fifth, the response-length analysis was exploratory and may have been influenced by model-specific communication styles and variation in question complexity; the observed associations should therefore not be interpreted as causal. Sixth, the study evaluated responses to standardized public-facing questions rather than individualized clinical scenarios and did not assess advice tailored to laboratory findings, medications, hepatic or renal function, treatment stage, comorbidities, or patient preferences. Finally, because proprietary model architectures, training data, retrieval processes, and safety-alignment mechanisms are not fully observable, the mechanisms underlying differences between models could not be determined directly.

Conclusion

In the primary standardized evaluation of five publicly accessible chatbots, most first-round responses to public-facing HCC nutrition questions were classified as safe, although clinically relevant unsafe or insufficiently qualified responses remained. Accuracy, empathy, information quality, and readability varied across systems, whereas the overall binary safety comparison did not, indicating that safety alone does not capture the broader quality of patient-facing chatbot performance. Response length showed only weak associations with the evaluated outcomes, suggesting that greater verbosity does not necessarily translate into better-quality information. Repeated live-query assessment further demonstrated generally acceptable short-term numerical stability across several non-safety outcomes under the study conditions, although these findings should not be interpreted as establishing long-term

temporal reproducibility of chatbot outputs. Overall, the findings support a cautious adjunctive role for LLMs in patient education rather than independent use for nutrition decisions that depend on liver function, treatment status, comorbidities, medication exposure, or red-flag symptoms. Future studies should evaluate performance across longer time intervals, model updates, languages, and user settings; incorporate patients and caregivers into comprehension and usability testing; and determine whether clinician-reviewed LLM-generated information can improve access to understandable nutrition guidance without compromising safety.

Declarations

Ethics approval and consent to participate

This study was approved by the Institutional Ethics Committee (Approval No. KY25067, 2025). The study evaluated large language model–generated responses to standardized hypothetical questions and did not involve the recruitment of patients, clinical interventions, medical records, or identifiable patient information. Five non-author clinical experts participated in the independent evaluation of chatbot responses. Two additional non-author experts provided methodological guidance and qualitative safety review, respectively. Informed consent was obtained from the external experts for their involvement in the study, and permission to identify them by name in the manuscript was obtained. As no patients or members of the public were recruited, patient informed consent was not applicable.

Privacy and data protection

All queries were hypothetical and standardized. No identifiable patient information, clinical records, laboratory data, imaging information, or personal health information were used during chatbot interaction. Data collection and analysis were conducted without disclosure of private or confidential information.

Permission and licensing

The study used investigator-developed questions and chatbot responses generated through publicly accessible web interfaces. No proprietary clinical datasets, copyrighted question banks, or restricted-access patient materials were used. The use of chatbot outputs was limited to academic evaluation and reporting purposes in accordance with applicable platform policies.

Protocol and registration

No prospective protocol registration was performed for this comparative chatbot evaluation with a

repeated-query stability component. The study design, question set development, query strategy, reference mapping, evaluation criteria, and statistical analysis plan are described in the Methods section and supplementary materials.

Data availability

The final question set, question-screening record, reference-answer mapping table, complete first-round response records, and summary analysis tables are provided in the supplementary materials.

Funding

This study received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors. The authors received no financial or material support from any chatbot provider or commercial artificial intelligence company.

Competing interests

The authors declare that they have no competing interests.

Author contributions

Xiaoli Qian, Zhaole Gong, and Xiaodong Cao contributed to the conception and design of the study. Xiaoli Qian and Qin Wu independently screened and refined the candidate questions. Xiaoli Qian and Weidong Wang developed the reference-answer mapping table and evaluation framework. Xiaoli Qian and Ling Miao jointly conducted primary-round chatbot response generation and output collection. Xiaoli Qian, Ling Miao, and Zhaole Gong jointly conducted the second-round live queries and response collection. Zhaole Gong and Xiaoli Qian performed the readability assessment and data recording, while Zhaole Gong was responsible for statistical data analysis. Xiaodong Cao supervised data organization, methodological refinement, and quality control and served as the corresponding author. Xiaoli Qian drafted the initial manuscript. All authors contributed to data interpretation, critically revised the manuscript, and reviewed and approved the final version. Xiaoli Qian and Ling Miao contributed equally to this work and share first authorship.

Acknowledgments

The authors thank Professor Zhishui Wu for his methodological guidance and constructive comments during the development of this study. The authors also thank Fang Fang for expert review of the qualitative safety classification.

Supplementary materials

Supplementary Appendix S1: Completed CHART checklist. Supplementary Appendix S2: Question

screening process and final question set. Supplementary Appendix S3: Reference-answer mapping table. Supplementary Appendix S4: Complete primary-round chatbot outputs. Supplementary Appendix S5: Readability formulas. Supplementary Tables S1–S6: model identifiers and access details; pairwise safety comparisons; significant pairwise comparisons for non-binary outcomes; significance-filtered pairwise safety summary; all pairwise comparisons for non-binary outcomes; and qualitative classification of responses coded as potentially unsafe or misleading.

Declaration of generative AI and AI-assisted technologies

During the preparation of this manuscript, generative AI and AI-assisted tools were used solely to improve language quality, including grammar, clarity, and readability, and to refine the visual presentation of the study workflow diagram (Figure 1). These editorial applications were independent of the chatbot systems evaluated in the study. AI-assisted tools were not used to modify the original chatbot responses, assign safety or quality ratings, perform statistical analyses, interpret findings, or make scientific decisions. All AI-assisted revisions were critically reviewed and verified by the authors, who take full responsibility for the accuracy, integrity, and final content of the manuscript.

References

1. Mauro E, de Castro T, Zeitlhoefler M, et al. Hepatocellular carcinoma: Epidemiology, diagnosis and treatment. JHEP Rep. 2025;7(12):101571. doi:10.1016/j.jhepr.2025.101571.

2. European Association for the Study of the Liver. EASL Clinical Practice Guidelines on the management of hepatocellular carcinoma. J Hepatol. 2025;82(2):315-374. doi:10.1016/j.jhep.2024.08.028.

3. Singal AK, Wong RJ, Dasarathy S, et al. ACG Clinical Guideline: Malnutrition and Nutritional Recommendations in Liver Disease. Am J Gastroenterol. 2025;120(5):950-972. doi:10.14309/ajg.0000000000003379.

4. Muscaritoli M, Arends J, Bachmann P, et al. ESPEN practical guideline: Clinical nutrition in cancer. Clin Nutr. 2021;40(5):2898-2913. doi:10.1016/j.clnu.2021.02.005.

5. Lai JC, Tandon P, Bernal W, et al. Malnutrition, frailty, and sarcopenia in patients with cirrhosis: 2021 Practice Guidance by the American Association for the Study of Liver Diseases. Hepatology. 2021;74(3):1611-1644. doi:10.1002/hep.32049.

6. Busch F, Hoffmann L, Rueger C, et al. Current applications and challenges in large language models for patient care: a systematic review. Commun Med (Lond). 2025;5(1):26.

674 doi:10.1038/s43856-024-00717-2.
675 7. Pan A, Musheyev D, Bockelman D, Loeb S, Kabarriti AE. Assessment of artificial intelligence
676 chatbot responses to top searched queries about cancer. JAMA Oncol. 2023;9(10):1437-1440.
677 doi:10.1001/jamaoncol.2023.2947.
678 8. Chen S, Kann BH, Foote MB, et al. Use of artificial intelligence chatbots for cancer treatment
679 information. JAMA Oncol. 2023;9(10):1459-1462. doi:10.1001/jamaoncol.2023.2954.
680 9. Chen D, Parsa R, Hope A, et al. Physician and artificial intelligence chatbot responses to cancer
681 questions from social media. JAMA Oncol. 2024;10(7):956-960. doi:10.1001/jamaoncol.2024.0836.
682 10. Chen D, Chauhan K, Parsa R, et al. Patient perceptions of empathy in physician and artificial
683 intelligence chatbot responses to patient questions about cancer. NPJ Digit Med. 2025;8(1):275.
684 doi:10.1038/s41746-025-01671-6.
685 11. Johri S, Jeong J, Tran BA, et al. An evaluation framework for clinical use of large language models
686 in patient interaction tasks. Nat Med. 2025;31(1):77-86. doi:10.1038/s41591-024-03328-5.
687 12. Yeo YH, Samaan JS, Ng WH, et al. Assessing the performance of ChatGPT in answering questions
688 regarding cirrhosis and hepatocellular carcinoma. Clin Mol Hepatol. 2023;29(3):721-732.
689 doi:10.3350/cmh.2023.0089.
690 13. Cao JJ, Kwon DH, Ghaziani TT, et al. Large language models' responses to liver cancer
691 surveillance, diagnosis, and management questions: accuracy, reliability, readability. Abdom Radiol
692 (NY). 2024;49(12):4286-4294. doi:10.1007/s00261-024-04501-7.
693 14. The CHART Collaborative. Reporting guideline for chatbot health advice studies: the Chatbot
694 Assessment Reporting Tool (CHART) statement. Br J Surg. 2025;112(8):znaf142.
695 doi:10.1093/bjs/znaf142.
696 15. Ayers JW, Poliak A, Dredze M, et al. Comparing physician and artificial intelligence chatbot
697 responses to patient questions posted to a public social media forum. JAMA Intern Med.
698 2023;183(6):589-596. doi:10.1001/jamainternmed.2023.1838.
699 16. Charnock D, Shepperd S, Needham G, Gann R. DISCERN: An instrument for judging the quality
700 of written consumer health information on treatment choices. J Epidemiol Community Health.
701 1999;53(2):105-111. doi:10.1136/jech.53.2.105.
702 17. Moult B, Franck LS, Brady H. Ensuring quality information for patients: development and
703 preliminary validation of a new instrument to improve the quality of written health care information.

Health Expect. 2004;7(2):165-175. doi:10.1111/j.1369-7625.2004.00273.x.
18. 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.
19. 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.
20. Smith EA, Senter RJ. Automated Readability Index. AMRL-TR-66-220. Wright-Patterson Air Force Base, OH: Aerospace Medical Research Laboratories; 1967.
21. Flesch R. A new readability yardstick. J Appl Psychol. 1948;32(3):221-233. doi:10.1037/h0057532.
22. Kincaid JP, Fishburne RP Jr, Rogers RL, Chissom BS. Derivation of new readability formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy enlisted personnel. Research Branch Report 8-75. Millington, TN: Naval Technical Training Command; 1975.
23. Gunning R. The Technique of Clear Writing. New York, NY: McGraw-Hill; 1952.
24. Coleman M, Liau TL. A computer readability formula designed for machine scoring. J Appl Psychol. 1975;60(2):283-284. doi:10.1037/h0076540.
25. McLaughlin GH. SMOG grading—a new readability formula. J Read. 1969;12(8):639-646.
26. Will J, Gupta M, Zaretsky J, Dowlath A, Testa P, Feldman J. Enhancing the readability of online patient education materials using large language models: cross-sectional study. J Med Internet Res. 2025;27:e69955. doi:10.2196/69955.
27. Fleiss JL. Measuring nominal scale agreement among many evaluators. Psychol Bull. 1971;76(5):378-382. doi:10.1037/h0031619.
28. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J Chiropr Med. 2016;15(2):155-163. doi:10.1016/j.jcm.2016.02.012.
29. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Series B Stat Methodol. 1995;57(1):289-300. doi:10.1111/j.2517-6161.1995.tb02031.x.
30. Gwet KL. Computing inter-rater reliability and its variance in the presence of high agreement. Br J Math Stat Psychol. 2008;61(Pt 1):29-48. doi:10.1348/000711006X126600.
31. Musheyev D, Pan A, Gross P, et al. Readability and information quality in cancer information from

734 a free vs paid chatbot. JAMA Netw Open. 2024;7(7):e2422275.
735 doi:10.1001/jamanetworkopen.2024.22275.
736 32. Siu AHY, Gibson DP, Chiu C, et al. ChatGPT as a patient education tool in colorectal cancer—an
737 in-depth assessment of efficacy, quality and readability. Colorectal Dis. 2025;27(1):e17267.
738 doi:10.1111/codi.17267.
739 33. Franc JM, Hertelendy AJ, Cheng L, Hata R, Verde M. Accuracy of a commercial large language
740 model (ChatGPT) to perform disaster triage of simulated patients using the Simple Triage and Rapid
741 Treatment (START) protocol: gage repeatability and reproducibility study. J Med Internet Res.
742 2024;26:e55648. doi:10.2196/55648.
743 34. Wang Y, Chen Y, Sheng J. Assessing ChatGPT as a medical consultation assistant for chronic
744 hepatitis B: cross-language study of English and Chinese. JMIR Med Inform. 2024;12:e56426.
745 doi:10.2196/56426.
746 35. Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. N Engl
747 J Med. 2023;388(13):1233-1239. doi:10.1056/NEJMsr2214184.
748 36. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language
749 models in medicine. Nat Med. 2023;29(8):1930-1940. doi:10.1038/s41591-023-02448-8.
750 37. Haupt CE, Marks M. AI-generated medical advice—GPT and beyond. JAMA.
751 2023;329(16):1349-1350. doi:10.1001/jama.2023.5321.
752 38. Clusmann J, Kolbinger FR, Muti HS, et al. The future landscape of large language models in
753 medicine. Commun Med (Lond). 2023;3(1):141. doi:10.1038/s43856-023-00370-1.
754 39. Wornow M, Xu Y, Thapa R, et al. The shaky foundations of large language models and foundation
755 models for electronic health records. NPJ Digit Med. 2023;6(1):135. doi:10.1038/s41746-023-00879-8.

Supplementary Files

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

- SupplementaryAppendixS2.docx
- SupplementaryAppendixS3.docx
- SupplementaryAppendixS4.docx
- SupplementaryAppendixS5.docx
- SupplementaryTablesS1S6.docx