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.

Quality and Readability of Artificial Intelligence Chatbot Responses to Patient Questions About Exocrine Pancreatic Insufficiency: A Comparative Study

Donghong Wang
Wuzhou Medical College
Baihui Chen
Fangchenggang Traditional Chinese Medicine Hospital
Jialin Wang
771881434@qq.com

Fangchenggang Traditional Chinese Medicine Hospital
Yalu Chen
The First Affiliated Hospital of Guangxi University of Traditional Chinese Medicine
Xian Deng
Fangchenggang Traditional Chinese Medicine Hospital
Xiaoyi Huang
Fangchenggang Traditional Chinese Medicine Hospital
Qing Chen
Fangchenggang Traditional Chinese Medicine Hospital

Research Article

Keywords: artificial intelligence, large language models, exocrine pancreatic insufficiency, patient education, information quality, readability

Posted Date: August 10th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10521656/v1

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

Additional Declarations: No competing interests reported.

Abstract

Objective

To compare the information quality, source transparency, educational value, and readability of responses generated by four widely used artificial intelligence chatbots to patient questions about EPI and determine whether the content met the sixth-grade reading level recommended for patient education materials.

Methods

A web-based cross-sectional comparative design was used. Terms of global interest over the previous 5 years were identified through Medical Subject Headings and Google Trends. After deduplication and assessment of applicability, 20 core questions were developed. Each question was entered into ChatGPT 5.6, Copilot, Gemini 3.6 Flash, and Perplexity using a standardized procedure. Response quality was assessed with DISCERN, EQIP, the JAMA benchmarks, and the Global Quality Score (GQS). Readability was assessed with the Automated Readability Index (ARI), Gunning Fog Index (GFI), Flesch-Kincaid Grade Level (FKGL), Coleman-Liau Index, Simple Measure of Gobbledygook (SMOG), and Flesch Reading Ease Score (FRES). Quality scores are presented as medians and interquartile ranges. Models were compared with the Kruskal-Wallis test, and ε^2 was reported as the effect size. Readability results were compared with sixth-grade thresholds.

Results

A total of 80 responses were obtained. No overall differences among the four models were statistically significant for DISCERN, EQIP, JAMA, or GQS scores, with P values of 0.187, 0.529, 0.392, and 0.467, respectively. Median DISCERN scores ranged from 35.00 to 38.50, indicating poor treatment information quality or scores near the boundary between poor and fair quality. Median EQIP scores ranged from 52.50 to 55.00, indicating good quality with minor problems. The median GQS was 4.00 for all models, indicating favorable organization and usefulness. The median JAMA score was 0 for all models, as source attribution, authorship information, disclosures, and update dates were generally absent. Median values for all grade-level readability measures were substantially higher than 6, and median FRES values ranged from 13.04 to 30.76, all below 80. Perplexity and Copilot were easier to read on most measures, whereas Gemini had the greatest linguistic complexity. None of the four models met the recommended reading level.

Conclusions

The four artificial intelligence chatbots generated well-structured and broadly usable information about EPI. Treatment information quality and source transparency remained limited, and linguistic complexity substantially exceeded the recommended level for general patient education materials. These responses are currently best used as supplementary information after professional verification. Developers should provide authoritative citations, indicate the currency of guidelines, generate plain-language content, and adapt responses to patients' health literacy.

Introduction

Exocrine pancreatic insufficiency (EPI) is a clinical syndrome of impaired digestion and absorption caused by pancreatic exocrine secretion and/or intraluminal pancreatic enzyme activity that is insufficient for normal nutrient digestion. Common manifestations include steatorrhea or diarrhea, abdominal bloating, increased flatulence, weight loss, fat-soluble vitamin deficiencies, and protein-energy malnutrition. Patients with chronic pancreatitis, pancreatic cancer, cystic fibrosis, or a history of pancreatic surgery are at high risk of EPI [1, 2]. EPI may also develop after acute pancreatitis. Previous meta-analyses have found persistent pancreatic exocrine dysfunction during recovery in some patients, with severe pancreatitis and alcohol-related etiology further increasing the risk[3]. Endoscopic retrograde cholangiopancreatography (ERCP) is widely used to diagnose and treat pancreaticobiliary disorders, and post-ERCP pancreatitis is one of its most common complications[4, 5]. Pancreatic exocrine dysfunction should therefore be considered when gastrointestinal symptoms or declining nutritional status persist during recovery.

The recognition and management of EPI require clinical symptom assessment, fecal elastase testing, evaluation of nutritional status, appropriate dosing and timing of pancreatic enzyme replacement therapy (PERT), dietary management, and monitoring of vitamins and other micronutrients[2, 6]. Because effective management depends heavily on patients' understanding of the disease and its treatment, they need accurate, complete, and comprehensible information. Previous reviews have shown that EPI can impair quality of life, that patients actively seek information about the disease and its treatment, and that some adjust their PERT dosage or discontinue treatment on their own[7]. Access to clear and reliable information may therefore influence treatment behavior, symptom control, and nutritional management.

Generative artificial intelligence and large language models have made chatbots such as ChatGPT, Copilot, Gemini, and Perplexity readily available for disease explanations and health advice. Comparative studies in maintenance hemodialysis and sexually transmitted infections have found differences among models in DISCERN, EQIP, JAMA, and GQS scores, although source transparency and readability remained common limitations[8, 9]. A synthesis of 42 studies found that response quality and readability rarely reached optimal levels together; higher-quality texts often remained written at advanced reading levels[10]. Specific prompts can improve the readability of artificial intelligence-generated patient education materials, but accuracy, completeness, personalization, and consistency still require human review [11, 12].

Standardized, multimodel comparisons of responses to common patient questions about EPI are lacking. The condition has nonspecific symptoms, a complex diagnostic pathway, and treatment that depends heavily on patient adherence. It is therefore useful to assess whether chatbots provide traceable, well-structured, and comprehensible information. This study was designed and reported according to the core principles of the Chatbot Assessment Reporting Tool (CHART) statement[13]. We compared the information quality, transparency, educational value, and readability of responses from four widely used artificial intelligence chatbots to EPI-related questions. We asked whether quality scores differed among models and whether the generated content met the sixth-grade reading level recommended for patient education materials.

Methods

Study Design and Setting

This web-based cross-sectional comparative evaluation was conducted in July 2026 through a publicly accessible internet environment in the Guangxi Zhuang Autonomous Region, China. The study language was English, and Google Trends was set to worldwide. The study material comprised text responses generated by four artificial intelligence chatbots to standardized patient questions about EPI. No patients were recruited, no clinical intervention was performed, and no identifiable personal information was collected.

Question Identification, Screening, and Finalization

The standardized term “Exocrine Pancreatic Insufficiency” was first confirmed in the Medical Subject Headings database. Top related queries and Rising related queries worldwide over the previous 5 years were then retrieved from Google Trends using this term and related expressions. Google Trends reports relative search interest, so data for low-volume searches may be unstable, and aggregation by topic and related query may omit clinically relevant content[14]. Candidate questions were therefore supplemented with relevant clinical guidelines and expert documents, primarily guidelines on the diagnosis and management of EPI, prevention of post-ERCP pancreatitis, and ERCP-related adverse events[1, 2, 6, 15, 16]. Patient perspectives were added to the information needs identified through Google Trends and clinical guidance. Questions of practical concern were collected and synthesized from the National Pancreas Foundation Pancreatitis Support Community on Inspire and the HealthUnlocked Pancreatitis Community managed by Mission:Cure. Two researchers independently screened content related to EPI, pancreatitis, and PERT using predefined search terms and eligibility criteria. Irrelevant, duplicate, highly similar, or poorly formulated items that could not be converted into clear patient questions were excluded. Disagreements were resolved through discussion and, when necessary, adjudication by a third researcher. The research team deidentified the included content, merged semantically redundant items, and standardized each item as a complete English patient question while preserving its original

meaning. The process yielded 50 initial candidate items. After screening, deduplication, and standardization, 20 patient-oriented questions were retained.

Selection of Artificial Intelligence Chatbots

ChatGPT 5.6, Microsoft Copilot, Google Gemini 3.6 Flash, and Perplexity were included. The models were selected based on public accessibility, user scale, differences in technological approaches, and frequent use in previous evaluations of medical information. All were accessed through their official web interfaces with the default settings available at the time of data collection.

Standardized Input and Data Collection

Relevant browser data were cleared before data collection, and personalization settings that could affect response content were disabled to minimize the influence of browsing history, contextual memory, and personalized settings. Each question was submitted once in a separate new conversation. No role assignment, professional identity, word limit, reading-grade requirement, or citation request was added. No follow-up question, regeneration, or manual correction was performed. Questions were entered into each model in a predetermined order, and the complete first response was retained. This procedure simulated direct questioning by the general public under default conditions. In total, 80 responses were collected, with 20 from each of the four models.

Information Quality Assessment

Four complementary instruments were used. DISCERN is a 16-item instrument that assesses the clarity and sources of treatment information, uncertainty in the evidence, treatment options, and their benefits and risks. Total scores range from 16 to 75: 63 to 75 is excellent, 51 to 62 good, 39 to 50 fair, 27 to 38 poor, and 16 to 26 very poor[17]. EQIP assesses the completeness, structure, language, and presentation of patient education materials. The 20-item version was used, and scores were converted to a 0 to 100 scale. Scores of 76 to 100 indicated excellent quality, 51 to 75 good quality with minor problems, 26 to 50 substantial quality problems, and 0 to 25 serious quality problems[18]. The GQS uses a 1 to 5 scale to assess overall quality, information flow, organization, and usefulness to patients; 1 indicates very poor quality and 5 excellent quality[19]. The JAMA benchmarks assess authorship, attribution, disclosure, and currency. Each item is scored 0 or 1, for a total score of 0 to 4 [20].

Two raters with clinical or research experience in gastrointestinal disorders independently scored the responses. Before formal scoring, 10% of the questions were randomly selected for pilot assessment and calibration of the rating criteria. Model names were removed from the response texts and replaced with random identifiers to reduce bias related to brand recognition. A third senior rater reviewed disagreements and determined the final score.

Readability Assessment

Before readability analysis, responses were converted to plain text. Bullet points, numbering, URLs, in-text citation markers, and separate reference lists were removed without changing sentence content, wording, or meaning. Six measures were calculated using Python 3.10.4 and textstat 0.7.13: the Automated Readability Index (ARI), Gunning Fog Index (GFI), Flesch-Kincaid Grade Level (FKGL), Coleman-Liau Index (CL), Simple Measure of Gobbledygook (SMOG), and Flesch Reading Ease Score (FRES). Higher ARI, GFI, FKGL, CL, and SMOG values indicate a higher required reading grade; higher FRES values indicate easier reading. Under commonly used standards for patient education materials, a response met a sixth-grade reading level when FRES was at least 80 and all other grade-level measures were below 6 [21].

Statistical Analysis

After assessment of distributional characteristics, continuous variables were summarized as medians and interquartile ranges. DISCERN, EQIP, JAMA, and GQS scores were compared among the four models with the Kruskal-Wallis test, and Kruskal-Wallis ε^2 was reported as the effect size. If an overall test reached statistical significance, pairwise comparisons were planned with Dunn's test and Holm adjustment for multiple comparisons. Because none of the overall quality tests reached statistical significance, no post hoc pairwise comparisons were performed. Readability measures for each model were compared with the prespecified sixth-grade thresholds using two-sided Wilcoxon signed-rank tests and interpreted with the medians and observed ranges. A two-sided P value < 0.05 was considered statistically significant.

Results

Core Questions and Dataset Characteristics

The question-selection process yielded 20 core questions covering the definition, symptoms, treatment, diet, malnutrition, fecal elastase testing, PERT administration, over-the-counter pancreatic enzymes, risks associated with acute and chronic pancreatitis, timing of pancreatic function assessment, management of missed doses, persistent symptoms, vitamin monitoring, and post-ERCP pancreatitis. Each model generated 20 responses, producing a dataset of 80 texts. Table 1 lists the questions.

Table 1
Twenty Core Patient Questions About Exocrine Pancreatic Insufficiency Included in the Study

	Question
1	What is exocrine pancreatic insufficiency?
2	What are the symptoms of exocrine pancreatic insufficiency?
3	How is exocrine pancreatic insufficiency treated?
4	What diet is recommended for patients with exocrine pancreatic insufficiency?
5	Can exocrine pancreatic insufficiency cause malabsorption and nutritional deficiencies?
6	What is the fecal elastase test, and how is it used to diagnose exocrine pancreatic insufficiency?
7	What is pancreatic enzyme replacement therapy?
8	How should pancreatic enzymes be taken?
9	Are over-the-counter pancreatic enzyme supplements equivalent to prescription pancreatic enzymes?
10	Can acute pancreatitis cause exocrine pancreatic insufficiency?
11	Why does chronic pancreatitis cause exocrine pancreatic insufficiency?
12	How can symptoms of exocrine pancreatic insufficiency be distinguished from symptoms of chronic pancreatitis?
13	Which patients with pancreatitis are at higher risk of developing exocrine pancreatic insufficiency?
14	When should pancreatic exocrine function be assessed after acute pancreatitis?
15	Can watery diarrhea affect the accuracy of the fecal elastase test?
16	Should pancreatic enzymes be taken with snacks as well as main meals?
17	What should a patient do after missing a dose of pancreatic enzymes?
18	What should patients do if symptoms persist despite pancreatic enzyme replacement therapy?
19	Do patients with exocrine pancreatic insufficiency need vitamin supplementation and nutritional monitoring?
20	Can post-ERCP pancreatitis lead to exocrine pancreatic insufficiency?

Information Quality Assessment Results

Kruskal-Wallis tests found no statistically significant differences among models in DISCERN ($P = 0.187$, $\varepsilon^2=0.024$), EQIP ($P = 0.529$, $\varepsilon^2=0.000$), JAMA ($P = 0.392$, $\varepsilon^2=0.000$), or GQS ($P = 0.467$, $\varepsilon^2=0.000$) scores. Table 2 reports the median, interquartile range, overall P value, and ε^2 effect size for each measure.

DISCERN had a very small effect size, whereas the effect sizes for the other measures were close to zero, indicating limited between-model differences.

Table 2
Information Quality Scores of the Four Artificial Intelligence Chatbots

Measure	ChatGPT	Copilot	Gemini	Perplexity	ε^2	P
DISCERN	37.00 (33.75, 39.25)	38.50 (34.00, 42.25)	37.00 (32.00, 43.50)	35.00 (33.75, 37.00)	0.024	0.187
EQIP	52.50 (48.75, 60.00)	55.00 (53.75, 60.00)	55.00 (45.00, 60.00)	55.00 (50.00, 61.25)	0.000	0.529
JAMA	0.00 (0.00, 0.00)	0.00 (0.00, 0.00)	0.00 (0.00, 0.00)	0.00 (0.00, 0.00)	0.000	0.392
GQS	4.00 (4.00, 4.00)	4.00 (3.75, 4.00)	4.00 (4.00, 4.00)	4.00 (3.00, 4.00)	0.000	0.467

Note: Values are presented as median (Q1, Q3). P values were obtained using the Kruskal-Wallis test, and ε^2 represents the Kruskal-Wallis effect size.

Median DISCERN scores were 37.00 for ChatGPT and Gemini and 35.00 for Perplexity, all within the poor-quality range. Copilot had a median score of 38.50, close to the boundary between poor and fair quality. Median EQIP scores ranged from 52.50 to 55.00, indicating good quality with minor problems. The median GQS was 4.00 for all models, reflecting favorable organization, information flow, and usefulness to patients. The median JAMA score was 0 for all models, indicating that most responses lacked identifiable authorship, clear attribution, disclosure information, and an update date. Figure 1 shows overlapping DISCERN and EQIP distributions across models. GQS scores were concentrated mainly between 3 and 4, whereas JAMA scores showed a pronounced floor effect.

Readability Assessment Results

Median values for all grade-level measures were substantially higher than 6, and median FRES values were substantially lower than 80. Table 3 summarizes the six readability measures for each model. Perplexity had the lowest median ARI, GFI, and CL values and the highest median FRES value. Copilot had the lowest median FKGL and SMOG values. Gemini had the highest median ARI, GFI, FKGL, CL, and SMOG values and the lowest median FRES value, indicating greater linguistic complexity.

Table 3
Readability Scores of the Four Artificial Intelligence Chatbots

Measure	ChatGPT	Copilot	Gemini	Perplexity	Sixth-grade threshold
ARI	16.94 (14.76, 18.49)	14.77 (13.33, 15.92)	18.95 (16.01, 20.60)	14.79 (11.98, 17.75)	< 6
GFI	18.27 (16.50, 19.69)	17.22 (13.97, 18.50)	19.68 (17.22, 22.05)	16.18 (12.90, 18.37)	< 6
FKGL	14.96 (13.80, 16.45)	13.47 (11.84, 14.70)	16.30 (14.12, 18.00)	13.93 (9.87, 15.56)	< 6
CL	18.45 (16.37, 19.48)	17.09 (15.54, 19.12)	18.94 (16.42, 20.02)	15.34 (12.54, 16.97)	< 6
SMOG	16.11 (14.84, 16.76)	13.52 (12.09, 15.53)	17.14 (15.41, 18.59)	14.93 (12.20, 16.04)	< 6
FRES	16.56 (5.78, 28.47)	21.50 (12.31, 32.26)	13.04 (5.64, 25.10)	30.76 (19.07, 48.30)	≥ 80

Note: Values are presented as median (Q1, Q3). A sixth-grade reading level was defined as ARI, GFI, FKGL, CL, and SMOG values below 6 and an FRES value of at least 80.

All grade-level measures exceeded the prespecified sixth-grade thresholds for every model.

FRES values for all models were also below the acceptable range. At the individual-response level, no response met the prespecified standard across all six measures. Figure 2 shows substantial dispersion across questions within each model. Perplexity and Copilot produced more concise responses to some questions, but even their lowest reading-grade values remained above the recommended standard.

Question-Level Differences in Scores

Figure 3 shows question-level quality and readability scores for the 20 questions across the four models. JAMA scores were close to 0 for almost all questions and models, indicating a consistent lack of source transparency. DISCERN and EQIP varied across questions, suggesting that question content influenced response quality. Complex questions involving differential interpretation, risk stratification, timing of follow-up, management of persistent symptoms, and nutritional monitoring generally produced higher grade-level scores. Gemini had greater reading complexity for several questions, whereas Perplexity and Copilot had lower scores for some questions. Considerable overlap among models remained.

Discussion

Principal Findings

We compared the quality and readability of responses generated by four widely used artificial intelligence chatbots to common patient questions about EPI. Three findings emerged. First, the models did not differ significantly in DISCERN, EQIP, JAMA, or GQS scores, and between-model effect sizes were very small. Second, the instruments separated distinct quality dimensions. EQIP and GQS indicated favorable structure, organization, and usefulness, whereas DISCERN placed treatment information quality in the poor range or near the boundary with fair quality, and JAMA showed limited source transparency. Third, all generated texts substantially exceeded the sixth-grade reading level. Perplexity and Copilot were easier to read on some measures, whereas Gemini was more complex, but these differences did not remove barriers to patient comprehension.

Users may therefore receive responses that are well structured, fluent, and apparently useful but still cannot verify the sources, assess the currency of the evidence, or understand technical terminology. The combination of relatively high GQS scores and JAMA scores close to 0 is a central finding. Surface usability and evidence traceability are distinct quality dimensions, and patients may trust fluent responses more than their source transparency warrants.

Comparison With Previous Studies and Interpretation of Between-Model Differences

Studies in other clinical settings have reported inconsistent model rankings. An inflammatory bowel disease study used DISCERN, EQIP, GQS, the JAMA benchmarks, and multiple readability measures and found significant differences among models[22]. In palliative care, Perplexity had the highest modified DISCERN and JAMA scores, whereas Gemini had the highest GQS score, showing that rankings may depend on the evaluation dimension[23]. A scoliosis patient education study found similar DISCERN scores despite differences in readability, and a study of chemotherapy-related cardiotoxicity found no model that consistently outperformed the others across patient education quality measures[24, 25].

We found no significant differences among the four models in DISCERN, EQIP, GQS, or JAMA scores. The composition of the questions, score distributions, and sample size may explain this result. Most questions addressed common topics, including the definition and symptoms of EPI, fecal elastase testing, PERT, and nutritional management. The models may therefore have generated similarly structured content from broadly comparable medical knowledge frameworks. GQS scores were concentrated at 4 and JAMA scores at 0, producing ceiling and floor effects that may have reduced sensitivity to small differences among models. The inclusion of only 20 questions may also have limited statistical power. The absence of statistically significant differences does not establish that the four models had identical information quality.

Systematic reviews have found substantial heterogeneity in medical chatbot studies, including question sources, query procedures, model versions, and assessment methods, with inconsistent performance and model rankings across studies[10, 26]. Repeated-query and cross-version studies have also shown that responses to the same question may vary with query time and model version[27]. Our single-time-point cross-sectional design represents performance only on the specific query dates and model

versions examined and cannot capture longitudinal changes during model updates. Future studies should repeat testing at multiple time points with identical questions and assessment criteria to evaluate stability and temporal trends.

Median EQIP scores for all four models fell within the 51 to 75 range despite the lack of significant between-model differences. Under the prespecified categories, the responses were acceptable as patient education materials but still had gaps in completeness, structure, and presentation[18]. EQIP assesses content and presentation, so a relatively high score does not establish that the text is easy for general patients to understand. In this study, readability measures substantially exceeded recommended levels, showing that favorable information quality can coexist with high reading difficulty. Previous syntheses have reported the same pattern[28]. DISCERN places greater weight on the reliability and decision-support value of treatment information, including evidence sources, treatment alternatives, benefits and risks, uncertainty, and shared decision-making[17]. The responses rarely explained the basis for PERT adjustment, further evaluation when treatment was ineffective, nutritional monitoring, or when professional care should be sought. These omissions may have contributed to the lower DISCERN scores. The JAMA benchmarks assess authorship, attribution, disclosure, and currency[20]. General-purpose chatbots usually omit these metadata in default responses, which explains the uniformly low JAMA scores. Because the JAMA benchmarks were developed for static online health information, their scores in generative conversations are best interpreted as proxy measures of source transparency. DISCERN, EQIP, GQS, and JAMA assess different dimensions: treatment information quality, patient education material quality, overall usability, and source transparency. The results characterize information quality, presentation, and transparency; they do not fully validate factual medical accuracy or guideline concordance.

Clinical Implications in the Context of EPI Information

Patient questions about EPI often prompt action. They include whether pancreatic enzymes should be taken with snacks, how to manage a missed dose, whether to increase the dose when symptoms persist, whether watery diarrhea affects fecal elastase results, and when to assess pancreatic exocrine function after acute pancreatitis. Patients with EPI actively seek information about the disease and its treatment, and some may adjust their PERT dosage on their own[29]. Our findings therefore have direct clinical implications. Although chatbot-generated responses were favorably rated for organization and usability, treatment information quality, source transparency, and readability remained limited. Clear structure and fluent language may increase patients' trust. When responses lack evidence sources, conditions of applicability, and explicit advice on when to seek medical care, patients may apply general information to individualized treatment decisions.

Chatbots may help explain basic EPI concepts, prepare questions for clinical consultations, and review health education provided by healthcare professionals. Previous reviews suggest that large language models may answer patient questions, explain complex medical information, and support patient-clinician communication, but limitations remain in accuracy, completeness, readability, and safety[30, 31]. Responses about PERT dosage or administration, interpretation of fecal elastase results, persistent

symptoms, deterioration in nutritional status, or complication risks should direct patients to a gastroenterologist or nutrition professional. Current guidelines state that EPI diagnosis should integrate high-risk underlying conditions, clinical symptoms, nutritional status, and appropriate pancreatic function testing[1]. PERT should be taken with food, and treatment response should be monitored through gastrointestinal symptoms, body weight, muscle status, and nutritional indicators such as fat-soluble vitamin levels[2]. Patients should not use chatbot responses that lack defined applicability, risk warnings, and professional verification to interpret test results or modify treatment regimens on their own.

Reasons for Inadequate Readability and Directions for Improvement

Most grade-level measures ranged from approximately grades 13 to 19, and FRES values fell within the very difficult range. EPI information inherently contains multisyllabic terms such as “exocrine pancreatic insufficiency,” “fecal elastase,” “steatorrhea,” “malabsorption,” “fat-soluble vitamins,” and “pancreatic enzyme replacement therapy.” Model responses also often used long sentences, nested conditions, parallel lists, and academic explanations, which increased syntactic and lexical complexity. Readability formulas rely mainly on sentence length, word length, and syllable count. Medical terminology therefore raises the scores, and general readers may struggle even when the text is logically structured.

Prompts that request a lower reading grade, shorter sentences, common words, and stepwise explanations of technical terms can improve the readability of generated content[12, 28]. This study used default prompts to reflect how patients ask questions directly. Requiring patients to request “please explain this in simple language” is unlikely to address differences in health literacy. Developers could automatically activate a plain-language mode for medical questions, define technical terms in parentheses at first mention, limit sentence and paragraph length, and provide both a “plain-language version” and a “professional version.” Actionable guidance on PERT administration should follow a clear sequence and use conditional statements, for example, “Take it with the first bite of food,” “Take it with snacks as prescribed,” and “Do not double the next dose after a single missed dose.”

Practical Implications

Patients may use chatbots to understand basic concepts, organize questions for clinical visits, and review information already explained by healthcare professionals. For questions about diagnosis, medication dosage, test results, or persistent symptoms, they should verify the information against guidelines or materials from healthcare institutions and consult a qualified professional. They should also check whether responses provide authoritative sources, publication or update dates, and conditions of applicability.

Healthcare professionals should ask whether patients use chatbots to obtain information about EPI, identify misconceptions arising from online responses, and provide reviewed patient education materials at an appropriate reading level. Clinical teams can develop standardized answers to common questions

and explain the timing of PERT administration, nutritional monitoring, and indications for follow-up during discharge education or outpatient review.

Developers should build retrieval-augmented generation systems based on authoritative guidelines and peer-reviewed evidence, display verifiable citations and evidence dates, set safety boundaries for high-risk questions, and test patient comprehension directly. Quality assessment should include factual accuracy, guideline concordance, actionability, risk of omission, and consistency across multi-turn conversations as well as fluency.

Strengths and Limitations

The study focused on EPI, a condition with clear patient education needs that has not been systematically evaluated in large language model research. The questions were derived from standardized subject terminology and global search trends and covered diagnosis, treatment, diet, nutrition, pancreatitis, and ERCP-related scenarios. DISCERN, EQIP, JAMA, GQS, and six readability measures assessed treatment information quality, presentation, usefulness, transparency, and linguistic complexity. The standardized conversation procedure and question-level heat maps improved reproducibility and interpretability.

This study has several limitations. It included only four models and 20 English-language questions, so the findings may not generalize to other models, languages, or diseases. The models are continuously updated, and the results provide only a snapshot of performance in July 2026; later versions may differ. Each question was submitted once, and variation from repeated generation, temperature settings, account differences, or multi-turn conversations was not evaluated. Google Trends does not report users' age, educational level, disease status, or actual search intent, and the question set could not capture all individualized information needs. DISCERN, EQIP, GQS, and JAMA were developed mainly for static health materials and have limited capacity to assess dynamic dialogue or interactive clarification. Readability formulas measure surface-level text complexity and cannot replace direct testing of patient comprehension. The study did not assess clinical accuracy, completeness, guideline concordance, actionability, or safety triage, so quality and readability scores alone cannot establish whether responses are suitable for independent treatment guidance. Interrater reliability data and complete model build identifiers should be verified against the original scoring records and platform interfaces before formal submission.

Future Research

Future studies could use the CHART framework for preregistered, repeated-generation evaluations at multiple time points and record complete model identifiers, access methods, internet connectivity, system prompts, and data-collection times. Question sets should include expressions used by patients with different health literacy levels, typographical errors, vague symptoms, and multi-turn follow-up questions. Patients and clinical experts should jointly evaluate comprehension, trust, actionability, and risk. For EPI, a fact-checking checklist could use AGA and European guidelines as reference standards, with particular attention to interpretation of fecal elastase results, PERT dosage and timing, nutritional

monitoring, follow-up after acute pancreatitis, and management of persistent symptoms. Further studies could compare default responses with plain-language prompts, retrieval-augmented generation, and fine-tuning with specialist knowledge bases.

Conclusion

ChatGPT 5.6, Copilot, Gemini 3.6 Flash, and Perplexity produced well-organized and well-presented responses to common patient questions about EPI, but treatment information quality was low and source transparency remained limited. The four models did not differ significantly on the quality measures, and every model produced language well above the sixth-grade reading level recommended for patient education materials. Perplexity and Copilot performed better on some readability measures, but their responses still did not meet the reading needs of general patients. Artificial intelligence chatbots may be used as adjuncts to EPI education after professional verification. Clinical usability may improve with verifiable citations, clear indication of guideline currency, plain-language constraints, and safety escalation mechanisms.

Declarations

Acknowledgements

The authors have no acknowledgements to declare.

Funding:

This study was supported by the project “Research and Application Demonstration of Key Technologies for Standardized Post-ERCP Management and the Prevention and Treatment of Complications Using Integrated Traditional Chinese and Western Medicine” (Grant No. Guike FN2600640105).

Data Availability Statement:

The datasets generated and/or analyzed during the current study are available from the corresponding author on reasonable request.

Conflict of Interest:

All authors have completed the ICMJE uniform disclosure form. The authors declare that they have no conflicts of interest to disclose.

Author Contributions

Conceptualization: D.H.W., B.H.C., and J.L.W.; Methodology: D.H.W., B.H.C., Y.L.C., and J.L.W.; Formal analysis: D.H.W., B.H.C., and X.D.; Investigation: D.H.W., B.H.C., Y.L.C., X.D., X.Y.H., and Q.C.; Data curation: D.H.W., B.H.C., Y.L.C., and X.D.; Validation: X.Y.H., Q.C., and J.L.W.; Visualization: D.H.W. and B.H.C.; Supervision: J.L.W.; Project administration: X.Y.H; Funding acquisition: X.Y.H.; Writing, original draft:

D.H.W. and B.H.C.; Writing, review and editing: D.H.W., B.H.C., J.L.W., Y.L.C., X.D., X.Y.H., and Q.C. All authors reviewed and approved the final manuscript.

D.H.W. and B.H.C. contributed equally to this work and share first authorship.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable. This manuscript does not contain any identifiable personal data, images, or quotations from individual persons.

References

1. Dominguez-Muñoz JE, et al. European guidelines for the diagnosis and treatment of pancreatic exocrine insufficiency: UEG, EPC, EDS, ESPEN, ESPGHAN, ESDO, and ESPCG evidence-based recommendations. United Eur Gastroenterol J. 2025;13(1):125–72.
2. Whitcomb DC, Buchner AM, Forsmark CE. AGA Clinical Practice Update on the Epidemiology, Evaluation, and Management of Exocrine Pancreatic Insufficiency: Expert Review. Gastroenterology. 2023;165(5):1292–301.
3. Eperjesi O, et al. Risk factors for pancreatic exocrine insufficiency after acute pancreatitis: A systematic review and meta-analysis. Pancreatology. 2026;26(3):404–11.
4. Mukai S, et al. Clinical Practice Guidelines for post-ERCP pancreatitis 2023. Dig Endosc. 2025;37(6):573–87.
5. Akshintala VS et al. Incidence, severity, and mortality of post-ERCP pancreatitis: an updated systematic review and meta-analysis of 145 randomized controlled trials. Gastrointest Endosc, 2023. 98(1): p. 1–6.e12.
6. Phillips ME et al. Consensus for the management of pancreatic exocrine insufficiency: UK practical guidelines. BMJ Open Gastroenterol, 2021. 8(1).
7. Barkin JA, Delk TB, Powell VJ. Symptoms, burden, and unmet needs of patients living with exocrine pancreatic insufficiency: a narrative review of the patient experience. BMC Gastroenterol. 2024;24(1):101.
8. Cao J, et al. The reliability and readability of large language models in answering patient questions on maintenance hemodialysis: A comparative study. Digit Health. 2026;12:20552076261435836.
9. Yıldız HA, Söğütdelen E. AI Chatbots as Sources of STD Information: A Study on Reliability and Readability. J Med Syst. 2025;49(1):43.
10. Whittaker P, Sun M. Quality and readability of chatbot responses to patient questions: A systematic cross-sectional meta-synthesis. Health Inf J. 2025;31(4):14604582251388879.

11. Nasra M, et al. Can artificial intelligence improve patient educational material readability? A systematic review and narrative synthesis. Intern Med J. 2025;55(1):20–34.
12. Will J, et al. Enhancing the Readability of Online Patient Education Materials Using Large Language Models: Cross-Sectional Study. J Med Internet Res. 2025;27:e69955.
13. Collaborative C. Reporting guideline for chatbot health advice studies: the Chatbot Assessment Reporting Tool (CHART) statement. BMJ Med. 2025;4(1):e001632.
14. Rovetta A. Google trends in infodemiology: Methodological steps to avoid irreproducible results and invalid conclusions. Int J Med Inf. 2024;190:105563.
15. Buxbaum JL, et al. American Society for Gastrointestinal Endoscopy guideline on post-ERCP pancreatitis prevention strategies: summary and recommendations. Gastrointest Endosc. 2023;97(2):153–62.
16. Dumonceau JM, et al. ERCP-related adverse events: European Society of Gastrointestinal Endoscopy (ESGE) Guideline. Endoscopy. 2020;52(2):127–49.
17. Charnock D, et al. DISCERN: an instrument for judging the quality of written consumer health information on treatment choices. J Epidemiol Community Health. 1999;53(2):105–11.
18. Moult B, Franck LS, Brady H. Ensuring quality information for patients: development and preliminary validation of a new instrument to improve the quality of written health care information. Health Expect. 2004;7(2):165–75.
19. Bernard A, et al. A systematic review of patient inflammatory bowel disease information resources on the World Wide Web. Am J Gastroenterol. 2007;102(9):2070–7.
20. 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):1244–5.
21. Washington GN, et al. Is the Message Clear? Evaluation of Readability Levels for Cleft Lip, Cleft Palate, and Craniofacial Websites. Cleft Palate Craniofac J. 2023;60(12):1619–24.
22. Wang X, et al. Benchmarking public large language model responses to patient-facing inflammatory bowel disease questions: informational quality, transparency proxies, and readability. Front Public Health. 2026;14:1810358.
23. Hancı V, et al. Assessment of readability, reliability, and quality of ChatGPT®, BARD®, Gemini®, Copilot®, Perplexity® responses on palliative care. Med (Baltim). 2024;103(33):e39305.
24. Zhao M, et al. Evaluating the readability and quality of AI-generated scoliosis education materials: a comparative analysis of five language models. Sci Rep. 2025;15(1):35454.
25. Stephenson-Moe CA, et al. Assessing the quality and readability of patient education materials on chemotherapy cardiotoxicity from artificial intelligence chatbots: An observational cross-sectional study. Med (Baltim). 2025;104(15):e42135.
26. Wei Q, et al. Evaluation of ChatGPT-generated medical responses: A systematic review and meta-analysis. J Biomed Inf. 2024;151:104620.

27. Goodman RS, et al. Accuracy and Reliability of Chatbot Responses to Physician Questions. JAMA Netw Open. 2023;6(10):e2336483.
28. Ensari E, Onder ENA, Ertan P. A Comparative Assessment of Large Language Models in Pediatric Dialysis: Reliability, Quality and Readability. Ther Apher Dial. 2025;29(5):739–46.
29. Lewis DM, Shahid A. Survey of Pancreatic Enzyme Replacement Therapy Dosing Experiences in Adults with Exocrine Pancreatic Insufficiency. Healthc (Basel), 2023. 11(16).
30. Aydin S, et al. Large language models in patient education: a scoping review of applications in medicine. Front Med (Lausanne). 2024;11:1477898.
31. Busch F, et al. Current applications and challenges in large language models for patient care: a systematic review. Commun Med (Lond). 2025;5(1):26.

Figures

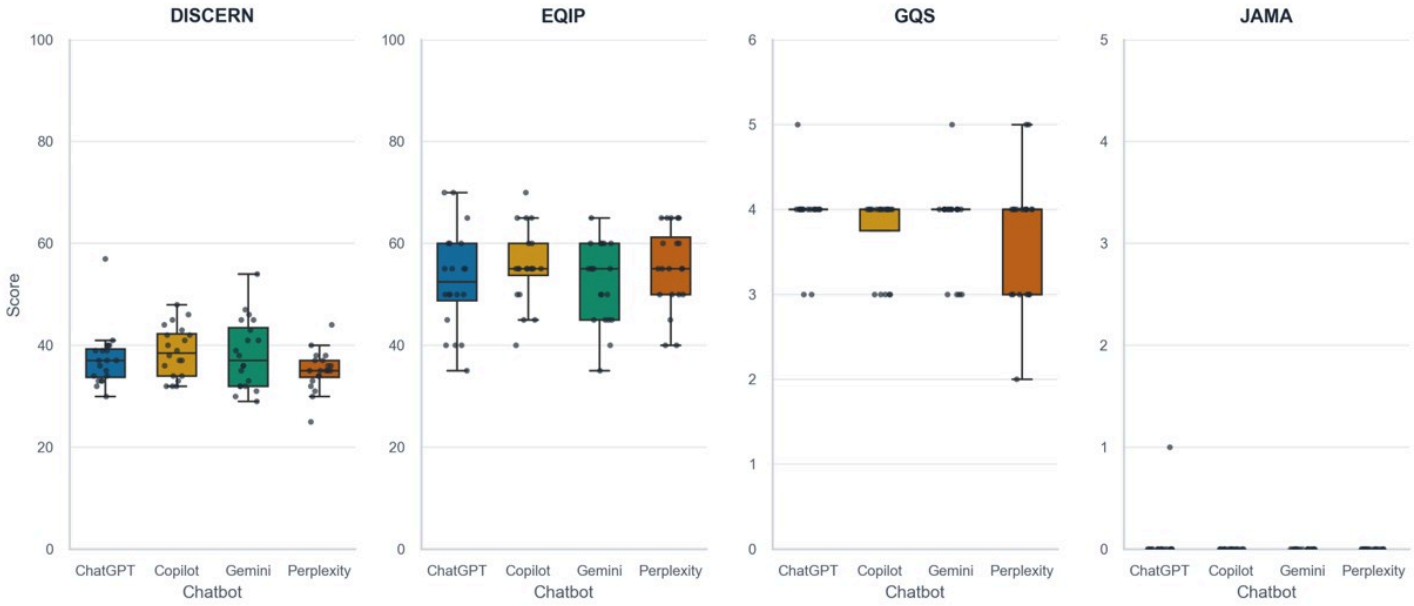


Figure 1

Distribution of Information Quality Scores Across the Four Artificial Intelligence Chatbots. Box plots show medians, interquartile ranges, outliers, and individual question-level scores.

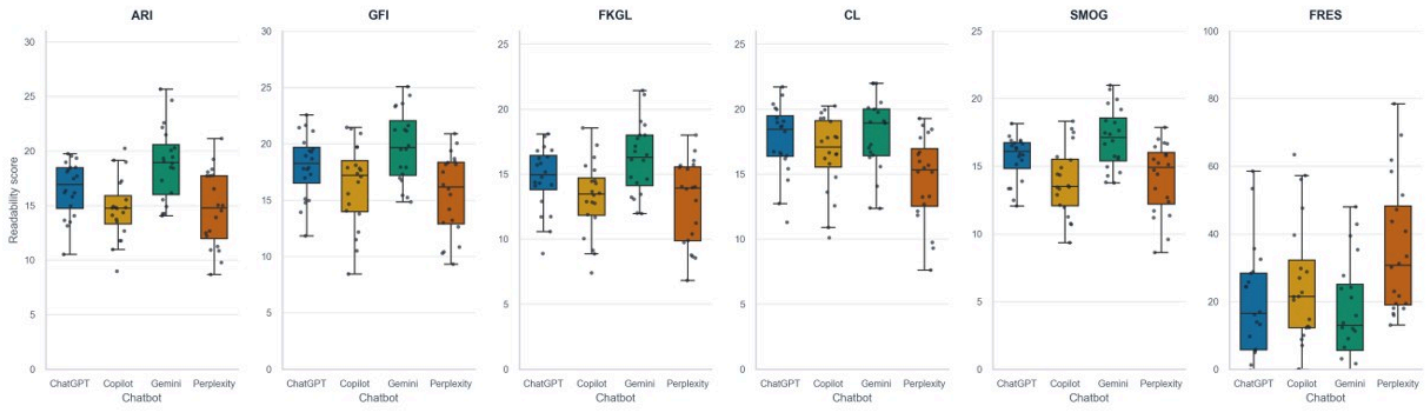


Figure 2

Distribution of Readability Measures Across the Four Artificial Intelligence Chatbots. Higher ARI, GFI, FKGL, CL, and SMOG values indicate greater reading difficulty, whereas higher FRES values indicate easier text.

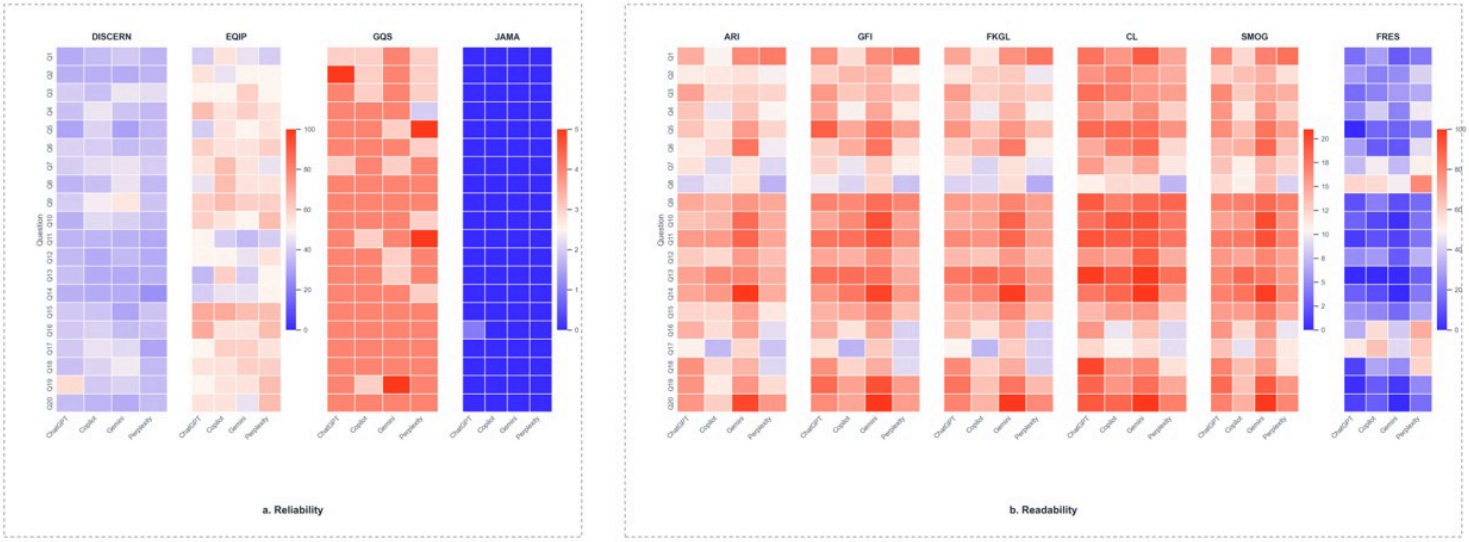


Figure 3

Question-Level Heat Maps of Information Quality and Readability. The left panel shows DISCERN, EQIP, GQS, and JAMA scores; the right panel shows ARI, GFI, FKGL, CL, SMOG, and FRES scores.