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.

Large Language Model–Assisted Personalized Patient Education for Chronic Atrophic Gastritis: A Blinded Paired Clinical Informatics Evaluation

Xinqiao Chu
Guang'anmen Hospital, China Academy of Chinese Medical Sciences
Wangyu Ye
Guang'anmen Hospital, China Academy of Chinese Medical Sciences
Yanjiao Zhang
Guang'anmen Hospital, China Academy of Chinese Medical Sciences
Wenwen Li
Guang’ anmen Hospital Baoding, China Academy of Chinese Medical Sciences
Wenying Liu
Guang’ anmen Hospital Baoding, China Academy of Chinese Medical Sciences
Zijian Li
Guang’ anmen Hospital Baoding, China Academy of Chinese Medical Sciences
Jianhai Niu
Guang’ anmen Hospital Baoding, China Academy of Chinese Medical Sciences
Shaoli Wang
Guang'anmen Hospital, China Academy of Chinese Medical Sciences
Weidong Li
Guang'anmen Hospital, China Academy of Chinese Medical Sciences
Muqing Zhang
The Fourth Affiliated Hospital of Hebei Medical University
Yixin Zhang
Hebei University of Chinese Medicine
Ji-zheng Ma
Guang'anmen Hospital, China Academy of Chinese Medical Sciences
Zhen Liu
drliuzhenky@163.com

Guang'anmen Hospital, China Academy of Chinese Medical Sciences

Article

Keywords: Chronic atrophic gastritis, Large language models, Artificial intelligence, Medical informatics, Patient education, Clinical decision support, Health communication

Posted Date: November 5th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11390227/v1

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

Additional Declarations: No competing interests reported.

Abstract

Background

Chronic atrophic gastritis (CAG) is a gastric precancerous condition requiring individualized communication regarding endoscopic and histopathological findings, Helicobacter pylori status, cancer risk, treatment, and surveillance. Large language models (LLMs) may facilitate individualized patient education, but evidence evaluating their performance within clinically matched and blinded assessment frameworks remains limited.

Objective

To evaluate the quality of large language model–assisted personalized patient education materials for chronic atrophic gastritis using a blinded clinical informatics framework and to compare their performance with clinician-generated materials.

Methods

In this single-center, blinded, paired comparative study, 50 patients with CAG were used to generate 200 patient education materials: one each from GPT-6.1 Sol, Qwen3.8 Flash, and two gastroenterologists. All sources received identical deidentified clinical information and standardized generation constraints. Three independent gastroenterology experts evaluated all materials using the 19-item Chronic Atrophic Gastritis Patient Education Quality Scoring System (CAG-PEQS), which assessed medical accuracy, personalization, comprehensiveness, and safety (overall score range, 0–95). Objective linguistic and structural characteristics were analyzed separately. Between-source comparisons were performed using Friedman tests followed by Holm-adjusted Wilcoxon signed-rank tests, and inter-rater reliability was assessed using intraclass correlation coefficients.

Results

Overall CAG-PEQS scores differed significantly among Qwen3.8 Flash, GPT-6.1 Sol, and the clinician reference (Friedman test, $P < 0.001$). Qwen3.8 Flash and GPT-6.1 Sol achieved high and comparable overall scores, with median scores of 82.67 (IQR 81.17–83.67) and 82.50 (IQR 81.42–83.58), respectively, whereas the clinician reference showed a lower median score of 52.58 (IQR 49.04–55.04). Both LLM-generated materials demonstrated higher scores than the clinician reference across all four CAG-PEQS domains, including medical accuracy, personalization, comprehensiveness, and safety. Inter-rater reliability for the overall CAG-PEQS score was high (ICC = 0.917, 95% CI 0.902–0.929). Objective linguistic analyses showed that both LLMs generated longer materials than the clinician reference, with median Chinese character counts of 405.5 for Qwen3.8 Flash, 395.0 for GPT-6.1 Sol, and 293.0 for the clinician reference.

Conclusions

Under standardized case-based conditions, LLM-generated CAG patient education materials achieved higher expert-rated quality scores than clinician-generated reference materials. These findings suggest that LLMs may have potential as clinician-supervised tools for drafting structured patient education materials. However, the observed differences in expert ratings do not establish superiority in real-world physician–patient communication or improvements in patient comprehension and clinical outcomes. Further studies incorporating patient-centered assessments, validated readability measures, and real-world clinical evaluation are needed to establish their clinical utility.

Introduction

Chronic atrophic gastritis (CAG) is a preneoplastic gastric condition characterized histologically by the loss of appropriate gastric glands, with or without metaplastic replacement, most commonly in association with chronic *Helicobacter pylori* (*H pylori*) infection or autoimmunity [1]. Its clinical importance extends beyond symptom control because the extent and severity of atrophy and the presence of intestinal metaplasia (IM) or dysplasia influence subsequent gastric cancer risk and surveillance strategies; current guidance therefore supports risk-stratified endoscopic follow-up for patients with gastric precancerous conditions [2]. *H pylori* assessment and appropriate eradication management constitute another central component of long-term care, while patients with histologic intestinal metaplasia or marked endoscopic atrophy may remain at increased risk of gastric cancer even after successful eradication [3, 4]. Consequently, effective CAG education requires more than providing general information about gastritis. Patients need understandable explanations of endoscopic and histopathologic findings, balanced communication about cancer risk, appropriate interpretation of *H pylori* status, and clear guidance regarding treatment adherence, lifestyle modification, follow-up, and alarm symptoms. Terms such as “atrophy,” “intestinal metaplasia” and “dysplasia” may otherwise be misunderstood or incorrectly equated with cancer, potentially resulting in either unnecessary anxiety or inadequate attention to surveillance.

Traditionally, these educational tasks are undertaken by clinicians during outpatient encounters. However, CAG education is inherently individualized: the information that should be emphasized may differ according to histologic severity, IM or low-grade dysplasia, *H pylori* status, current symptoms, previous treatment, risk factors, and available follow-up information. Clinicians must therefore translate heterogeneous clinical data into patient-friendly language while avoiding unsupported assumptions, excessive risk amplification, inappropriate treatment recommendations, and omission of clinically important follow-up information. These requirements are difficult to satisfy consistently within time-constrained clinical encounters. Recent work in chronic gastritis has also explored AI-agent-based approaches to support postdiagnosis health management [5]. A scalable approach capable of producing standardized yet case-specific educational content could therefore complement, rather than replace, clinician-delivered education.

Large language models (LLMs) have emerged as a potential tool for patient-facing medical communication because they can synthesize contextual information and generate natural-language

responses at scale. Comparative studies have evaluated the performance of LLMs in health consultation and patient education in areas such as urolithiasis, and LLM-assisted approaches have also been evaluated for responding to real-world patient messages [6, 7]. More recently, a randomized controlled trial demonstrated that LLM-generated patient-oriented discharge summaries could improve patient activation compared with standard discharge summaries, providing early evidence that generative artificial intelligence (AI) may affect patient-centered outcomes beyond conventional question-answering benchmarks [8]. Nevertheless, fluent and apparently personalized output does not necessarily indicate clinically reliable communication. Expert consensus has identified misinformation, inaccurate communication, privacy concerns, and overreliance as important risks of LLM use in health care, while broader ethical analyses have emphasized accountability, bias, transparency, and the need for appropriate clinical oversight [9, 10]. Model-specific variability represents an additional concern: a recent comparative study of ChatGPT-4o and DeepSeek-R1 identified model-specific differences in medical misinformation within patient-education outputs [11]. These limitations are particularly relevant to CAG, where unsupported statements regarding cancer progression, *H pylori* treatment, medication, or surveillance could lead to anxiety, inappropriate self-management, or delayed medical care.

Despite increasing interest in LLM-based patient education, evidence specifically addressing personalized education for histologically confirmed CAG remains limited. Previous work in chronic gastritis has primarily evaluated question-answering or AI-agent-based health management rather than a controlled comparison of educational materials generated from the same individual patient records. In other chronic diseases, blinded comparisons between LLM-generated and clinician-created personalized educational materials have demonstrated the feasibility of structured expert assessment, but these findings cannot be assumed to generalize to the distinctive risk-communication and surveillance requirements of CAG [12]. We therefore conducted a blinded, paired comparative study in which two contemporary LLMs and two gastroenterologists independently generated patient education materials from identical deidentified CAG case information under standardized task constraints. The primary objective was to compare the overall quality of LLM-generated and clinician-generated materials using the prespecified Chronic Atrophic Gastritis Patient Education Quality Scoring System (CAG-PEQS), a study-specific structured expert rating framework encompassing medical accuracy, personalization, comprehensiveness, and safety. Secondary objectives were to compare performance across the four CAG-PEQS domains—medical accuracy, personalization, comprehensiveness, and safety—and to characterize the objective linguistic and structural characteristics of the Chinese-language CAG education materials.

Methods

Study Design

This single-center, paired comparative clinical informatics study was conducted at Guang'anmen Hospital, China Academy of Chinese Medical Sciences. Fifty eligible patients with CAG were selected from clinical records between July 1 and July 7, 2026. All included cases had available histopathological

information, and direct personal identifiers were removed before study processing. The overall study workflow, including clinical information preparation, educational material generation, and blinded expert evaluation, is illustrated in Fig. 1.

For each matched clinical case, four independent patient education materials were generated: one by GPT-6.1 Sol, one by Qwen3.8 Flash, and one each by two gastroenterologists, resulting in 200 educational materials derived from 50 matched CAG cases.

To ensure comparability across generation sources, the two LLMs and the two physicians received identical deidentified clinical information and were assigned the same patient education task under standardized output constraints. Depending on availability within the clinical record, case information included demographic characteristics, presenting symptoms, upper gastrointestinal endoscopic findings, histopathological findings, Helicobacter pylori (H pylori) status, physician-prescribed treatments, follow-up recommendations, relevant risk factors, and documented patient concerns. Information not available in the source record was retained as missing and was not inferred, supplemented, or fabricated.

The two physician generators were associate chief physicians in gastroenterology with approximately 10 years of clinical experience. Both physicians routinely participated in the diagnosis and management of CAG-related conditions. Each independently generated patient education materials for all 50 cases and had no access to materials generated by either LLM or by the other physician.

Figure 1. Study design and workflow for generation and blinded evaluation of personalized CAG patient education materials.

LLM Generation and Prompt Framework

LLM-based generation was performed on August 4, 2026, through the OpenRouter Chat Completions API using GPT-6.1 Sol and Qwen3.8 Flash. Each clinical case was submitted as an independent API request containing the same predefined system prompt and a case-specific user message. No conversational history, image input, external retrieval, or web search was used during generation.

For GPT-6.1 Sol, temperature, top-p, top-k, and seed parameters were not explicitly specified, and max_completion_tokens was set to 2500. For Qwen3.8 Flash, temperature was set to 0.2, while top-p, top-k, and seed parameters were not explicitly specified, and max_tokens was set to 2500. Context compression was disabled. Provider fallback mechanisms, when applicable, could only route requests to another endpoint serving the same requested model and could not substitute a different model. All 100 planned LLM-case generation requests were completed successfully.

Both LLMs received the same CAG-specific system prompt structured according to a Purpose–Reference/Evidence Boundary–Input–Output–Safety (P-R-I-O-S) framework. The prompt required all patient-specific statements to be derived exclusively from the corresponding clinical information and prohibited fabrication of diagnoses, examination findings, medications, treatment courses, or physician

recommendations. The models were also instructed not to modify established treatment plans, recommend new medications or treatment durations, or infer unavailable H pylori results. When H pylori status was unavailable, the models were instructed to explicitly state that the available information was insufficient to determine infection status.

The P-R-I-O-S framework was designed to improve the reproducibility, safety, and transparency of clinical information transformation. The original Chinese prompt, English translation, model identifiers, API pathway, generation parameters, prompt provenance, request-integrity records, and material-level generation logs are provided in Supplementary Material S1.

Standardized Educational Structure

To ensure consistency across generation sources, both LLMs and physicians were required to generate materials using the same eight-section structure covering:

1. interpretation of examination findings;
2. explanation of CAG;
3. issues requiring current attention;
4. treatment purpose and adherence;
5. dietary and lifestyle recommendations;
6. follow-up planning;
7. alarm symptoms;
8. common misconceptions.

All generators were instructed to limit each complete educational material to no more than 500 Chinese characters. Original outputs were retained without subsequent editing, rewriting, correction, or manual truncation.

Expert Evaluation and Blinded Assessment

Three independent gastroenterology experts evaluated all generated materials. All reviewers were associate chief physicians in gastroenterology with approximately 5 years of clinical experience and had not participated in patient care or material generation for the study cases.

Before formal assessment, reviewers completed a calibration exercise using three non-study CAG cases. During calibration, reviewers discussed the definitions and behavioral anchors of the CAG Patient Education Quality Scoring System (CAG-PEQS) until a shared interpretation was achieved. Reviewers specifically agreed that the S3 medical-boundary item should assess functional safety, including appropriate clinician referral, follow-up guidance, specialist consultation, and urgent-care recommendations, rather than the simple presence of a generic disclaimer. Therefore, absence of a statement that the material could not replace a physician was not penalized independently. Calibration

cases were excluded from formal analysis, and all subsequent ratings were completed independently without discussion.

For source concealment, each educational material was assigned an anonymized study code, and all source-identifying information was removed. Reviewers received the corresponding deidentified clinical case together with the anonymized educational material. Materials were presented in randomized order, and reviewers were blinded to whether each material was generated by GPT-6.1 Sol, Qwen3.8 Flash, physician 1, or physician 2.

Because all four materials for each patient were generated from the same underlying clinical information, each clinical case was considered the matched analytical unit for statistical comparisons.

Outcomes

Two complementary outcome frameworks were evaluated separately: objective linguistic and structural characteristics of the generated patient education materials and expert-rated clinical quality using CAG-PEQS.

Objective Linguistic and Structural Characteristics

Objective characteristics of the original Chinese-language materials were quantified computationally without rewriting, editing, or translating the source texts. The prespecified analysis compared objective linguistic and structural characteristics among LLM-generated materials and expert-authored reference materials. To characterize linguistic and structural properties while minimizing redundant measurements, the analysis included nine objective metrics: Chinese character count, unique Chinese character count, Chinese sentence count, mean Chinese characters per sentence, maximum Chinese characters per sentence, Chinese character diversity ratio, lexical richness, mean noun sense count, and mean verb sense count. These metrics represented multiple dimensions of the analyzed materials, including text length, lexical diversity, sentence structure, vocabulary richness, and semantic complexity. Because these computational metrics do not constitute a validated Chinese-language patient readability instrument, they were interpreted as objective linguistic and structural characteristics rather than direct measures of patient comprehension, health literacy, or readability. Objective linguistic analyses were exploratory and were conducted separately from the CAG-PEQS ratings.

CAG-PEQS Expert Evaluation

The CAG-PEQS was a prespecified, study-specific expert rating framework developed for blinded assessment of CAG patient education materials. Item content and scoring anchors were informed by major guidance on atrophic gastritis, gastric precancerous conditions, *H pylori* management, and gastric cancer surveillance, including the American Gastroenterological Association Clinical Practice Update on atrophic gastritis, the MAPS II guideline, the Maastricht VI/Florence consensus report, the Guidelines for

Diagnosis and Treatment of Chronic Gastritis in China (2022, Shanghai), the Chinese Expert Consensus on the Management of Gastric Precancerous Conditions and Lesions (2020), and the National Health Commission's Gastric Cancer Screening and Early Diagnosis and Treatment Program (2024) [1–3, 13–15]. These sources informed the domains addressing disease definition, interpretation of atrophy, intestinal metaplasia and dysplasia, *H pylori* management, treatment explanation, surveillance, gastric cancer risk communication, personalization, comprehensiveness, and safety. The framework and its behavioral anchors were finalized before formal scoring.

The CAG-PEQS comprises 19 items across four domains: medical accuracy (6 items), personalization (6 items), comprehensiveness (4 items), and safety (3 items). Each item was rated on a 0–5 scale using predefined behavioral anchors, with higher scores indicating better performance. Domain scores were calculated by summing the corresponding item scores, yielding maximum scores of 30 for medical accuracy, 30 for personalization, 20 for comprehensiveness, and 15 for safety. The overall CAG-PEQS score was calculated as the sum of all 19 item scores, with a possible range of 0–95. The overall score was prespecified as the primary outcome, the four domain scores as secondary outcomes, and individual item scores as exploratory outcomes. Detailed item definitions, scoring anchors, and the guidance basis are provided in Supplementary Material S2.

For each education material, ratings from the three independent reviewers were averaged to obtain material-level overall and domain CAG-PEQS scores. For the prespecified primary source comparison, the scores of the two physician-generated materials were further averaged within each case to create a clinician reference, yielding three matched comparison groups: GPT-6.1 Sol, Qwen3.8 Flash, and the clinician reference. The three prespecified pairwise comparisons were GPT-6.1 Sol versus Qwen3.8 Flash, GPT-6.1 Sol versus the clinician reference, and Qwen3.8 Flash versus the clinician reference. In an additional prespecified analysis, GPT-6.1 Sol and Qwen3.8 Flash scores were averaged within each case to generate an AI-composite score, whereas the two physician scores were averaged to generate a clinician-composite score.

Statistical Analysis

Continuous variables were summarized as mean (SD) or median (IQR), as appropriate, and categorical variables as counts and percentages. All between-source analyses were performed at the case level to account for the paired study design.

Overall CAG-PEQS scores for GPT-6.1 Sol, Qwen3.8 Flash, and the clinician reference were compared using the Friedman test. Prespecified pairwise comparisons were performed using two-sided Wilcoxon signed-rank tests with Holm adjustment for multiple comparisons. Kendall's W and rank-biserial correlation were reported as effect-size measures for the global and pairwise comparisons, respectively. The same approach was applied to the four CAG-PEQS domain scores. AI-composite and clinician-composite scores were compared using a two-sided Wilcoxon signed-rank test.

Objective linguistic and structural characteristics were analyzed separately using the same paired nonparametric framework where appropriate. Inter-rater reliability among the three expert reviewers was assessed using intraclass correlation coefficients (ICCs) with 95% confidence intervals based on a two-way random-effects, absolute-agreement, average-measures model.

No missing values were imputed. All tests were two-sided, and *P* < *.05* was considered statistically significant after multiplicity adjustment where applicable. Statistical analyses were performed using Python 3.13.5 (Python Software Foundation).

Ethical Considerations

The study was approved by the Ethics Committee of Guang'anmen Hospital, China Academy of Chinese Medical Sciences (approval No. 2026-108-KY). Informed consent was obtained from all participants before their participation in the study. All study procedures were performed in accordance with the Declaration of Helsinki and applicable institutional and national guidelines and regulations. Only deidentified clinical information was used for the generation and evaluation of patient education materials. No AI-generated content was used for clinical decision-making or provided directly to patients.

Results

Objective Linguistic and Structural Characteristics

Complete paired objective linguistic and structural data were available for 50 cases. Significant differences were observed among OpenAI GPT-6.1 Sol, Qwen3.8 Flash, and the Expert Reference across the retained objective metrics. Qwen3.8 Flash generated slightly longer materials than GPT-6.1 Sol, with median Chinese character counts of 405.5 (IQR 387.25-419.75) and 395.0 (IQR 387.25–404.0), respectively, whereas the Expert Reference materials were shorter, with a median Chinese character count of 293.0 (IQR 282.5-305.5). Qwen3.8 Flash also demonstrated higher mean Chinese characters per sentence and maximum sentence length compared with GPT-6.1 Sol and the Expert Reference, whereas the Expert Reference showed the highest Chinese-character diversity ratio (median 0.6026), followed by GPT-6.1 Sol (0.5821) and Qwen3.8 Flash (0.5450). Differences were also observed in advanced language features. GPT-6.1 Sol showed higher lexical richness and mean noun sense count than Qwen3.8 Flash, whereas the Expert Reference demonstrated higher mean noun and verb sense counts, indicating distinct semantic feature profiles among the three source groups. These findings describe objective linguistic and structural differences among source groups and were not interpreted as validated measures of Chinese-language readability, patient comprehension, or health literacy.

Table 1
Objective linguistic and structural characteristics of patient education materials

Metric	GPT-6.1 Sol, median (IQR)	Qwen3.8 Flash, median (IQR)	Clinician Reference, median (IQR)	Kendall's W	Global P
Chinese Character Count	395.00 (387.25–404.00)	405.50 (387.25-419.75)	293.00 (282.50-305.50)	0.761	< 0.001
Unique Chinese Character Count	230.00 (225.25–235.00)	220.00 (212.25-225.75)	172.50 (170.00-178.50)	0.870	< 0.001
Chinese Sentence Count	21.00 (20.00–22.00)	20.00 (19.00–21.00)	18.50 (18.00-18.50)	0.529	< 0.001
Mean Chinese Characters per Sentence	18.91 (18.40–19.40)	20.07 (18.73–21.24)	15.65 (15.11–16.92)	0.526	< 0.001
Maximum Chinese Characters per Sentence	40.00 (37.00-45.75)	47.50 (41.50-52.75)	41.50 (38.50–47.00)	0.064	0.0407
Chinese Character Diversity Ratio	0.58 (0.57–0.59)	0.55 (0.53–0.56)	0.60 (0.59–0.61)	0.521	< 0.001
Lexical Richness	7.32 (7.26–7.37)	7.26 (7.21–7.32)	6.92 (6.87–6.96)	0.788	< 0.001
Mean Noun Sense Count	3.67 (3.52–3.82)	3.16 (2.95–3.32)	3.77 (3.64–3.92)	0.670	< 0.001
Mean Verb Sense Count	6.90 (6.64–7.18)	6.75 (6.49–7.29)	8.05 (7.84–8.39)	0.693	< 0.001

Overall CAG-PEQS Score

Overall CAG-PEQS scores differed significantly among Qwen3.8 Flash, GPT-6.1 Sol, and the Expert Reference (Friedman test, P < 0.001). Qwen3.8 Flash achieved the highest overall score, with a median score of 82.67 (IQR 81.17–83.67), followed closely by GPT-6.1 Sol (median 82.50, IQR 81.42–83.58). No significant difference was observed between the two LLM-generated sources, whereas both LLM-generated materials achieved significantly higher scores than the Expert Reference.

Figure 2. Distribution of overall CAG-PEQS scores across generation sources. Horizontal boxplots show case-level overall CAG-PEQS scores (0–95) for GPT-6.1 Sol, Qwen3.8 Flash, and the clinician reference across 50 matched CAG cases. Each point represents one clinical case. Boxplots indicate median and interquartile range. For each material, ratings were averaged across three independent expert reviewers. The clinician reference was calculated as the within-case mean of the two physician-generated materials.

Domain-Specific and Item-Level CAG-PEQS Performance

Significant differences were observed across all four CAG-PEQS domains among Qwen3.8 Flash, GPT-6.1 Sol, and the Expert Reference (all Friedman tests P < 0.001). Both LLM-generated materials demonstrated consistently higher domain-level scores than the Expert Reference, while Qwen3.8 Flash and GPT-6.1 Sol showed comparable performance across the evaluated domains. For medical accuracy, GPT-6.1 Sol achieved a median score of 26.33/30 (IQR 25.67–26.67), followed by Qwen3.8 Flash (26.00/30, IQR 25.33–26.67), whereas the Expert Reference showed a lower score of 17.58/30 (IQR 16.38–18.50). For personalization, Qwen3.8 Flash and GPT-6.1 Sol achieved similar scores (26.33/30 and 26.00/30, respectively), both exceeding the Expert Reference (15.58/30). Similarly, both LLMs demonstrated higher comprehensiveness scores (Qwen3.8 Flash: 17.50/20; GPT-6.1 Sol: 17.67/20) compared with the Expert Reference (10.67/20). In the safety domain, Qwen3.8 Flash and GPT-6.1 Sol maintained high scores (13.33/15 and 13.00/15, respectively), while the Expert Reference showed a lower score (8.83/15). These findings indicate that LLM-generated materials consistently provided more complete, individualized, and safety-oriented patient education content under the standardized evaluation framework.

Table 2
Overall and domain-specific CAG-PEQS scores by generation source

Domain	Qwen3.8 Flash Median (IQR)	GPT-6.1 Sol Median (IQR)	Clinician Reference Median (IQR)	Kendall's W	*P*
Medical Accuracy (0–30)	26.00 (25.33–26.67)	26.33 (25.67–26.67)	17.58 (16.38–18.50)	0.766	< 0.001
Personalization (0–30)	26.33 (25.00–26.67)	26.00 (25.33–26.33)	15.58 (14.04–16.46)	0.754	< 0.001
Comprehensiveness (0–20)	17.50 (17.00–18.00)	17.67 (16.83–18.00)	10.67 (10.33–11.12)	0.772	< 0.001
Safety (0–15)	13.33 (12.67–13.67)	13.00 (12.33–13.33)	8.83 (8.33–9.17)	0.785	< 0.001

Note: Values are based on 50 matched cases. For each material, scores were averaged across three independent expert reviewers. The clinician reference was calculated as the within-case mean of the two physician-generated materials. Global differences were assessed using the Friedman test. Pairwise comparisons were performed using two-sided Wilcoxon signed-rank tests with Holm adjustment. Both LLMs differed significantly from the clinician reference (all adjusted $P < 0.001$), whereas no significant difference was observed between the two LLMs. W, Kendall's coefficient of concordance.

Item-level analyses demonstrated a consistent pattern across the 19 CAG-PEQS criteria (Fig. 3). Qwen3.8 Flash and GPT-6.1 Sol achieved relatively high and stable scores across most items, whereas the Expert Reference showed greater variability between individual criteria. Both LLM-generated sources performed particularly well in items related to disease explanation, interpretation of clinical findings, *H*

pylori management, follow-up recommendations, and safety boundaries. The largest differences between LLM-generated and expert-authored materials were observed in items requiring systematic integration of patient-specific information, including *H pylori*-related personalization, risk-factor matching, treatment safety, and individualized follow-up guidance. These results suggest that the advantage of LLM-generated materials was primarily associated with consistent organization and explicit coverage of predefined patient education requirements rather than differences in basic medical information accuracy alone.

Figure 3. Item-level CAG-PEQS performance across generation sources. Each cell shows the mean item score on the 0–5 CAG-PEQS scale across 50 matched CAG cases. For each material, ratings were first averaged across the three independent expert reviewers. The clinician reference was then calculated as the within-case mean of the two physician-generated materials. Columns are grouped into medical accuracy (A1-A6), personalization (P1-P6), comprehensiveness (C1-C4), and safety (S1-S3). A common color scale was applied to all cells, with cyan-blue indicating lower scores and coral red indicating higher scores. Full item definitions and scoring anchors are provided in Supplementary Material S2.

Inter-rater Reliability

Inter-rater reliability was high for the overall CAG-PEQS score and all four domains (Table 3). The average-measures ICC was 0.917 (95% CI 0.902–0.929) for the overall 0–95 score and ranged from 0.821 for safety to 0.952 for comprehensiveness across the four domains.

Table 3
Inter-rater reliability of CAG-PEQS overall and domain scores

CAG-PEQS outcome	ICC	95% CI
Overall CAG-PEQS (0–95)	0.917	0.902–0.929
Medical accuracy (0–30)	0.875	0.856–0.891
Personalization (0–30)	0.926	0.912–0.937
Comprehensiveness (0–20)	0.952	0.939–0.962
Safety (0–15)	0.821	0.797–0.840

Discussion

Principal Findings

In this blinded, paired comparative study, both GPT-6.1 Sol and Qwen3.8 Flash generated personalized CAG patient education materials that achieved higher expert-rated quality scores than the clinician reference under standardized case-based conditions. Overall CAG-PEQS scores differed significantly among the three sources, while the two LLMs demonstrated comparable performance across the overall

score and the four predefined domains of medical accuracy, personalization, comprehensiveness, and safety. These findings suggest that contemporary LLMs can consistently transform structured clinical information into patient-facing educational materials that align with expert-defined quality criteria.

The observed advantage of LLM-generated materials was not limited to factual accuracy but extended to systematic integration of patient-specific information, coverage of clinically relevant educational components, and maintenance of safety boundaries. However, these findings should be interpreted within the context of a structured written evaluation framework. Higher expert-rated scores indicate greater adherence to predefined educational requirements but do not necessarily imply superior clinical judgment, diagnostic reasoning, or effectiveness in real-world patient communication. Rather, these findings suggest that LLMs may have potential as clinician-supervised drafting tools for producing structured patient education materials, although their effects on clinical workflow efficiency and patient-centered outcomes require further evaluation.

Performance of LLMs in Structured Patient Education

Large language models have increasingly been explored as tools for patient education and medical communication because of their ability to synthesize complex information and generate patient-oriented explanations. Previous studies comparing LLM-generated medical information with healthcare professional responses have suggested that LLMs may achieve comparable or, in some contexts, higher ratings for information quality, although substantial heterogeneity exists across clinical domains, evaluation methods, and model types [16]. Similarly, studies evaluating AI-generated patient education materials have demonstrated that LLMs can produce structured educational content with acceptable accuracy, understandability, and actionability, while potentially reducing the time required for material preparation [17]. Broader reviews have further summarized the potential applications of LLMs in patient education, including medical information explanation, lifestyle guidance, and generation of patient-facing educational materials, while emphasizing persistent concerns regarding accuracy, readability, bias, and safety [18, 19].

Our study extends these previous observations by evaluating personalized CAG education materials generated from identical patient information under a blinded, paired comparative framework. Unlike general medical question-answering tasks, which primarily assess the correctness of isolated responses, our study required LLMs and clinicians to transform comprehensive clinical information-including endoscopic findings, histopathological results, *H pylori* status, treatment plans, and follow-up recommendations-into individualized educational materials. The comparable performance of GPT-6.1 Sol and Qwen3.8 Flash, together with their higher CAG-PEQS scores than the clinician reference, suggests that contemporary LLMs can provide consistent and structured first-draft educational materials when the generation task is explicitly defined.

The strong performance observed in the present study likely reflects the interaction between model capability and structured task design rather than model capability alone. Both LLMs received a disease-specific prompt framework requiring explicit coverage of examination interpretation, disease

explanation, patient-specific concerns, treatment adherence, lifestyle modification, follow-up planning, alarm symptoms, and common misconceptions while restricting unsupported medical inference. Previous studies have similarly shown that prompt engineering can substantially influence the quality of LLM-generated patient education materials. For example, structured prompting has been reported to improve readability and organization of colorectal cancer education materials [20]. while prompt strategies in retinal disease education and chronic disease education have demonstrated that modifying instructions can alter readability, accuracy, and comprehensiveness [21–23]. More recent work in Chinese-language patient education has also shown that prompt design can influence accuracy, helpfulness, and linguistic complexity of LLM-generated medical responses [24]. Therefore, the high CAG-PEQS performance observed in our study should be interpreted as the combined effect of contemporary LLM capabilities, disease-specific prompt engineering, and a structured evaluation framework.

The higher CAG-PEQS scores observed for LLM-generated materials should be interpreted within the context of a structured written evaluation rather than as evidence of superior clinical communication. LLM-generated responses often contain more explicit explanations and greater information coverage, which may align well with rubric-based assessments emphasizing completeness and predefined educational elements. Previous comparisons between LLM-generated and physician responses have similarly reported that LLM outputs may be rated as more detailed, clear, or comprehensive in certain settings, while also highlighting concerns regarding excessive length, occasional inaccuracies, and the continued need for clinician oversight [25, 26]. Conversely, studies comparing LLM-generated materials with professional resources have found no significant differences in information quality, understandability, actionability, or empathy, and professional resources may retain advantages in readability and practical communication [27]. Therefore, the higher scores observed in our study likely reflect the ability of LLMs to consistently organize and explicitly present information under predefined requirements, rather than superiority in clinical judgment, patient interaction, or individualized decision-making.

Model Consistency Across Contemporary LLMs and Ceiling Effects

The observed results may have been influenced by the standardized P-R-I-O-S prompting framework and structured evaluation criteria; however, the independent contribution of prompt design was not directly evaluated. Although previous studies have reported differences among LLMs across various medical domains, these differences have not been consistently directional and appear to depend on the specific model version, language environment, prompt design, and evaluation framework. For example, previous comparisons between ChatGPT and DeepSeek have demonstrated model-dependent variation in patient education quality, including differences in completeness, readability, and explainability, while other studies have reported different models achieving advantages in different clinical scenarios [28–32]. These findings indicate that LLM performance should be interpreted as task-specific rather than reflecting a universally superior model. In the present study, the similar CAG-PEQS performance of

Qwen3.8 Flash and GPT-6.1 Sol likely reflects the influence of the standardized P-R-I-O-S prompting framework and structured evaluation criteria, rather than the intrinsic superiority of either model.

An additional consideration is the potential ceiling effect of the CAG-PEQS framework when evaluating high-performing LLM-generated materials. Although both LLMs achieved high expert-rated scores, the concentration of scores within the upper range suggests that some items may have limited ability to discriminate between highly competent outputs. Ceiling effects have been recognized as an important consideration in the evaluation of measurement instruments because excessive clustering near the maximum score may reduce sensitivity for detecting meaningful differences among high-performing groups [33]. However, CAG-PEQS was developed as a study-specific expert evaluation framework rather than a fully validated psychometric instrument. Therefore, upper-range clustering should be interpreted as a limitation in discriminative resolution rather than definitive evidence of instrument failure. Future refinement of CAG-PEQS may incorporate additional criteria addressing information prioritization, conciseness, redundancy control, and patient-centered communication to better distinguish among advanced AI-generated educational materials.

Linguistic and Structural Trade-offs

The objective linguistic analyses revealed a trade-off between information coverage and textual conciseness in LLM-generated patient education materials. Compared with the clinician reference, both Qwen3.8 Flash and GPT-6.1 Sol generated longer materials, with median Chinese character counts of 405.5 and 395.0, respectively, compared with 293.0 for the clinician reference. In addition to greater text length, LLM-generated materials demonstrated differences in sentence structure and lexical features, suggesting that contemporary LLMs tend to produce more explicitly structured and information-dense educational content. These characteristics may contribute to higher scores within an expert-based evaluation framework that emphasizes completeness, personalization, and safety-related information coverage. However, greater textual detail does not necessarily translate into improved patient understanding or communication effectiveness.

The relationship between linguistic complexity and patient education quality requires careful interpretation. In the present study, objective linguistic metrics were designed to characterize structural and semantic features of the generated materials rather than directly measure readability or comprehension. The higher Chinese-character diversity ratio observed in the clinician reference compared with LLM-generated materials should not be interpreted as evidence of superior linguistic quality, because diversity metrics such as type-token measures are influenced by text length and may decrease as repeated linguistic units accumulate in longer documents[34]. Moreover, previous evaluations of Chinese-language LLM applications have demonstrated that models may produce responses with different levels of linguistic complexity, length, and readability despite achieving similar levels of accuracy or completeness [35, 36]. Therefore, linguistic complexity should be considered as an independent dimension of patient education quality rather than a surrogate marker of clinical usefulness.

An important observation was that both LLMs achieved high CAG-PEQS scores despite producing longer materials than clinician-generated content. This finding highlights a potential tension between rubric-based quality assessment and real-world patient communication requirements. Structured evaluation systems may reward explicit coverage of clinical information, whereas patients may benefit from concise explanations that prioritize the most relevant information for decision-making. Previous studies have emphasized that accuracy, comprehensiveness, readability, and patient-centered accessibility do not always improve simultaneously[37]. Prompt-based approaches may help balance these competing objectives by encouraging LLMs to preserve essential information while reducing unnecessary complexity. Future evaluations should therefore combine expert-rated clinical quality with validated Chinese-language readability measures and direct patient assessments to determine the optimal balance between completeness, brevity, and comprehensibility [38]. More broadly, written health information frequently exceeds recommended readability levels, and accessibility should be considered independently from factual completeness [39]. Future evaluations should therefore combine expert-rated clinical quality assessments with validated Chinese-language readability measures and direct patient assessments to determine the optimal balance between completeness, brevity, and comprehensibility.

Clinical Implications

Our findings suggest that LLMs may have potential utility as clinician-supervised drafting tools for individualized CAG patient education, rather than as autonomous sources of patient-facing guidance. However, their clinical feasibility and effectiveness remain to be established in real-world settings. Under standardized case-based conditions, both GPT-6.1 Sol and Qwen3.8 Flash generated educational materials with high expert-rated quality across medical accuracy, personalization, comprehensiveness, and safety domains, suggesting that contemporary LLMs may assist clinicians in producing more consistent and structured written educational content. A potential clinical workflow may involve generating an initial patient education draft from deidentified clinical information, followed by physician verification, prioritization, and refinement before patient delivery. Similar human-in-the-loop approaches have been proposed for integrating LLMs into clinical communication workflows, emphasizing clinician oversight for factual verification, contextual adaptation, and maintenance of accountability [40–42]. In this workflow, clinicians should remain responsible for confirming medical accuracy, removing unnecessary or excessive information, adapting explanations to individual health literacy levels, and ensuring that treatment and follow-up recommendations remain consistent with the actual clinical plan. Importantly, higher CAG-PEQS scores should be interpreted as improved alignment with predefined expert evaluation criteria rather than direct evidence of superior patient comprehension or clinical outcomes. Future implementation studies should evaluate patient acceptability, understanding, behavioral impact, and real-world integration of LLM-assisted patient education workflows.

Strengths and Limitations

This study has several strengths. First, the paired design ensured that GPT-6.1 Sol, Qwen3.8 Flash, and both gastroenterologists generated educational materials from identical case information, thereby

minimizing case-mix differences across generation sources. Second, all materials were evaluated under source-blinded conditions by three independent gastroenterology experts after a predefined calibration procedure, and inter-rater reliability was high. Third, the use of a prespecified, disease-specific CAG-PEQS allowed patient education quality to be assessed across multiple clinically relevant dimensions rather than accuracy alone. The study also incorporated objective linguistic and structural analyses, providing complementary information regarding how generated materials differed in textual organization and complexity. Finally, the complete generation protocol was preserved, including exact model identifiers, API access route, generation date, parameters, prompt provenance, request checks, and material-level run logs, which is particularly important for reproducibility in LLM research [43, 44].

Several limitations should also be acknowledged. First, this was a single-center study involving 50 CAG cases and two LLMs; therefore, the findings may not generalize to other populations, clinical settings, languages, model versions, or future model updates. Second, although CAG-PEQS was prespecified, behaviorally anchored, and demonstrated high inter-rater reliability, it remains a study-specific framework without independent psychometric validation. Reliability alone does not establish other measurement properties, including content, construct, or criterion validity [45]. Third, the generation prompt and CAG-PEQS were both developed around key requirements for CAG patient education. This construct alignment may have favored outputs that explicitly followed predefined educational structures, particularly for models optimized for structured content generation. Fourth, although source-identifying information was removed and materials were presented in randomized order, differences in language style, length, and structure may have allowed reviewers to infer the generation source in some cases; the effectiveness of source blinding was not formally assessed. Fifth, although both LLMs generated longer materials than the clinician reference, text length may provide greater opportunity to address rubric-defined elements, and the influence of information volume on expert-rated scores cannot be fully excluded. Sixth, the clinician reference consisted of written educational materials generated by two gastroenterologists under standardized task constraints. These materials may not fully represent routine physician–patient communication, which involves interactive clarification, clinical judgment, and adaptation to individual patient needs. Therefore, the observed differences in expert-rated scores should not be interpreted as evidence that LLMs outperform physicians in actual clinical communication. Seventh, the study did not include direct patient assessment of comprehension, acceptability, recall, or behavioral outcomes; therefore, higher expert-rated CAG-PEQS scores cannot be assumed to translate into greater patient benefit. This distinction is important because current LLM evaluations may not fully capture real-world clinical effectiveness or translational impact [46]. Finally, the objective linguistic measures used here characterize textual structure rather than validated Chinese-language readability, and the results should not be interpreted as direct evidence of patient comprehension. Furthermore, the study did not assess material preparation time, clinician workload, or the efficiency of LLM-assisted workflows in routine clinical practice. Therefore, potential improvements in workflow efficiency remain hypothetical.

Conclusion

In this single-center, blinded, paired comparative study, GPT-6.1 Sol and Qwen3.8 Flash generated individualized patient education materials for chronic atrophic gastritis that achieved higher expert-rated CAG-PEQS scores than clinician-generated reference materials under standardized case-based conditions. No statistically significant differences were observed between the two LLMs in overall or domain-specific scores, although equivalence between the models was not established. These findings suggest that LLMs may have potential as clinician-supervised tools for drafting structured patient education materials. However, higher expert-rated quality does not establish superior clinical communication, improved patient comprehension, or better clinical outcomes. Further studies incorporating patient-centered assessments, validated readability measures, and real-world clinical evaluation are needed to determine the clinical utility of LLM-assisted patient education.

Abbreviations

AI
artificial intelligence
API
application programming interface
CAG
chronic atrophic gastritis
CAG-PEQS
Chronic Atrophic Gastritis Patient Education Quality Scoring System
CI
confidence interval
ICC
intraclass correlation coefficient
IM
intestinal metaplasia
IQR
interquartile range
LLM
large language model
MAPS II
Management of Epithelial Precancerous Conditions and Lesions in the Stomach II
P-R-I-O-S
Purpose-Reference/Evidence Boundary-Input-Output-Safety
SD
standard deviation

Declarations

Acknowledgments

We gratefully acknowledge the assistance provided by ChatGPT (OpenAI) during the preparation of this manuscript.

Funding

This work was supported by the National Key Research and Development Program of China (Grant No. 2025YFC3512803); the Key Collaborative Research Project of the Technological Innovation Program of the China Academy of Chinese Medical Sciences (No. CI2023C015YL), and the 2025 China University Industry-University-Research Innovation Fund-Industry Association Smart Healthcare Innovation Special Project (No. 2025XH039).

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Data Availability

The datasets generated and/or analyzed during the current study are not publicly available because they contain information subject to patient privacy considerations and institutional data protection policies. Deidentified data may be made available from the corresponding author upon reasonable request, subject to approval by the relevant institutional ethics committee and applicable data protection requirements.

Authors' contributions

Xinqiao Chu contributed to study design, data analysis, visualization, and drafting of the manuscript. Wangyu Ye contributed to methodology, statistical analysis, and manuscript revision. Yanjiao Zhang contributed to case screening, data curation, and data verification. Wenwen Li, Wenying Liu, and Zijian Li were responsible for preparing and organizing the CAG materials for expert evaluation and coordinating the collection of rating data. Muqing Zhang and Ji-zheng Ma independently prepared the clinician-generated personalized CAG patient education materials. Jianhai Niu, Shaoli Wang, and Yixin Zhang served as independent expert reviewers and performed blinded CAG-PEQS assessments. Weidong Li contributed to clinical methodology, interpretation of the findings, and manuscript revision. Zhen Liu contributed to conceptualization, supervision, funding acquisition, and critical revision of the manuscript. All authors reviewed and approved the final manuscript.

References

1. Shah, S. C., Piazuelo, M. B., Kuipers, E. J., Li, D. & AGA Clinical Practice Update on the Diagnosis and Management of Atrophic Gastritis. *Expert Rev. Gastroenterol.* ;**161**(4):1325–32e7. doi: 10.1053/j.gastro.2021.06.078. (2021).
2. Pimentel-Nunes, P. et al. Management of epithelial precancerous conditions and lesions in the stomach (MAPS II): European Society of Gastrointestinal Endoscopy (ESGE), European Helicobacter and Microbiota Study Group (EHMSG), European Society of Pathology (ESP), and Sociedade Portuguesa de Endoscopia Digestiva (SPED) guideline update 2019. *Endoscopy* **51** (04), 365–388. 10.1055/a-0859-1883 (2019).
3. Malfertheiner, P. et al. Management of Helicobacter pylori infection: the Maastricht VI/Florence consensus report. *Gut* **71** (9), 1724–1762. 10.1136/gutjnl-2022-327745 (2022). PMID: 35944925.
4. Shichijo, S. et al. Histologic intestinal metaplasia and endoscopic atrophy are predictors of gastric cancer development after Helicobacter pylori eradication. *Gastrointest. Endosc.* **84** (4), 618–624. 10.1016/j.gie.2016.03.791 (2016).
5. Wang, S. & Ye, Q. Enhancing Large Language Models With AI Agents for Chronic Gastritis Management: Comprehensive Comparative Study. *JMIR Med. Inf.* **13**, e73857. 10.2196/73857 (2025). PMID: 41236204.
6. Song, H. et al. Evaluating the Performance of Different Large Language Models on Health Consultation and Patient Education in Urolithiasis. *J. Med. Syst.* **47** (1). 10.1007/s10916-023-02021-3 (2023).
7. Chen, S. et al. The effect of using a large language model to respond to patient messages. *Lancet Digit. Health*. **6** (6), e379–e81. 10.1016/s2589-7500(24)00060-8 (2024).
8. Rust, P. et al. Effects of large language model-generated, patient-oriented discharge summaries on patient activation: a single-centre, single-blind, randomised controlled trial in Germany. Lancet Digit Health. ;8(6):100991. PMID: 42150962. (2026). 10.1016/j.landig.2026.100991
9. Denecke, K., May, R. & LlmhealthGroup, Rivera Romero, O. Potential of Large Language Models in Health Care: Delphi Study. *J. Med. Internet Res.* **26**, e52399. 10.2196/52399 (2024). PMID: 38739445.
10. Li, H. et al. Ethics of large language models in medicine and medical research. *Lancet Digit. Health*. **5** (6), e333–e5. 10.1016/s2589-7500(23)00083-3 (2023).
11. Wang, G. et al. Detection of Medical Misinformation in Hemangioma Patient Education: Comparative Study of ChatGPT-4o and DeepSeek-R1 Large Language Models. *Jmir Ai*. **4**, e76372–e. 10.2196/76372 (2025).
12. Du, K. et al. Comparing Artificial Intelligence–Generated and Clinician-Created Personalized Self-Management Guidance for Patients With Knee Osteoarthritis: Blinded Observational Study. *J. Med. Internet. Res.* **27** 10.2196/67830 (2025).
13. Chinese Society of Gastroenterology CCGoCSoGCMA. Guidelines for diagnosis and treatment of chronic gastritis in China & Shanghai). J Dig Dis. 2023;24(3):150–80. PMID: 37245073. (2022). 10.1111/1751-2980.13193

14. National Health Commission of the People's Republic of China. Gastric Cancer Screening and Early Diagnosis and Treatment Program (2024 Edition). *Chin. J. Oncol.* **46** (10), 915–916. 10.3760/cma.j.cn112152-20240705-00276 (2024).
15. National Clinical Research Center for Digestive Diseases, National Early Gastrointestinal Cancer Prevention and Treatment Center Alliance. Helicobacter pylori Group of the Chinese Society of Gastroenterology, Chinese Society of Gastroenterology. Chinese Consensus on Management of Gastric Epithelial Precancerous Conditions and Lesions (2020). *Chin. J. Digestion*. **40**, 731–741. 10.13558/j.cnki.issn1672-3686.2024.008.002 (2020).
16. Jacobs, M. M. G., Oosterhoff, J. H. F., Agricola, R. & van der Weegen, W. Large language models versus healthcare professionals in providing medical information to patient questions: A systematic review. Int J Med Inform. ;209:106250. PMID: 41529641. (2026). 10.1016/j.ijmedinf.2025.106250
17. Jiang, X. et al. Evaluating the Effectiveness of Generative AI for the Creation of Patient Education Materials on Coronary Heart Disease: A Comparative Study. *JMIR Form. Res.* **9**, e78816. 10.2196/78816 (2025). PMID: 41270264.
18. Busch, F. et al. Current applications and challenges in large language models for patient care: a systematic review. Commun Med (Lond). ;5(1):26. PMID: 39838160. (2025). 10.1038/s43856-024-00717-2
19. Aydin, S., Karabacak, M., Vlachos, V. & Margetis, K. Large language models in patient education: a scoping review of applications in medicine. Front Med (Lausanne). ;11:1477898. PMID: 39534227. (2024). 10.3389/fmed.2024.1477898
20. Ellison, I. E. et al. De novo generation of colorectal patient educational materials using large language models: Prompt engineering key to improved readability. Surgery. ;180:109024. PMID: 39756334. (2025). 10.1016/j.surg.2024.109024
21. Jung, H., Oh, J., Stephenson, K. A. J., Joe, A. W. & Mammo, Z. N. Prompt engineering with ChatGPT3.5 and GPT4 to improve patient education on retinal diseases. *Can. J. Ophthalmol.* **60** (3), e375–e81. 10.1016/j.jcjo.2024.08.010 (2025). PMID: 39245293.
22. Will, J. et al. Enhancing the Readability of Online Patient Education Materials Using Large Language Models: Cross-Sectional Study. *J. Med. Internet Res.* **27**, e69955. 10.2196/69955 (2025). PMID: 40465378.
23. Shafau, F., Wahl, C., Kado, M. & Miedema, G. Evaluating Large Language Model Adherence to Targeted Fifth-Grade Readability Standards in Patient Education on Chronic Conditions. *Chronic Dis. Transl Med.* **12** (1), 73–74. 10.1002/cdt3.70031 (2026). PMID: 42022584.
24. Wang, Z. et al. Evaluating the efficacy of large language models in cardio-oncology patient education: a comparative analysis of accuracy, readability, and prompt engineering strategies. *Front. Artif. Intell.* **8**, 1693446. 10.3389/frai.2025.1693446 (2025). PMID: 41608034.
25. Motegi, M. et al. Comparison of physician and large language model chatbot responses to online ear, nose, and throat inquiries. Sci Rep. ;15(1):21346. PMID: 40596359. (2025). 10.1038/s41598-025-06769-1

26. Richter, S. et al. Impact of Large Language Model-Based AI Tools on Physician-Patient Communication: Systematic Review and Meta-Analysis. *J. Med. Internet Res.* **28**, e77307. 10.2196/77307 (2026). PMID: 42536971.
27. Alabdalhussein, A. et al. Large Language Models vs. Professional Resources for Post-Treatment Quality-of-Life Questions in Head and Neck Cancer: A Cross-Sectional Comparison. Curr Oncol. ;32(12). PMID: 41440196. (2025). 10.3390/curroncol32120668
28. Lau, J. Y. S., Gerald Sng, G. R., Cao, R. & Chen, J. A comparative study of ChatGPT and DeepSeek in spinal cord injury patient education: Can artificial intelligence speak spinal cord injury? *J. Spinal Cord Med.* **49** (3), 618–623. 10.1080/10790268.2025.2554013 (2026). PMID: 40938207.
29. Hu, X. et al. Evaluation of large language models in patient education for hyperthyroidism: A comparative study of chatgpt, gemini, and deepseek. *Endocrine* **90** (3), 1332–1338. 10.1007/s12020-025-04421-6 (2025). PMID: 40944884.
30. Liu, Y. et al. Assessing the Role of Large Language Models Between ChatGPT and DeepSeek in Asthma Education for Bilingual Individuals: Comparative Study. *JMIR Med. Inf.* **13**, e65365. 10.2196/65365 (2025). PMID: 40802989.
31. Cabuk Celik, N. & Altunel Kilinc, E. AI-generated patient education for ankylosing spondylitis: a comparative study of readability and quality. *Clin. Rheumatol.* **45** (3), 2003–2008. 10.1007/s10067-025-07771-8 (2026). PMID: 41390886.
32. Karagoz, B., Bayrak, H. C. & Kececi, T. Comparison of ChatGPT-5 and DeepSeek V3 for Artificial Intelligence-Assisted Patient Education in Foot and Ankle Disorders. J Am Podiatr Med Assoc. ;116(3). PMID: 42200996. (2026). 10.3390/japma116030033
33. Terwee, C. B. et al. Quality criteria were proposed for measurement properties of health status questionnaires. *J. Clin. Epidemiol.* **60** (1), 34–42. 10.1016/j.jclinepi.2006.03.012 (2007). PMID: 17161752.
34. Koizumi, R. & In'nami, Y. Effects of text length on lexical diversity measures: Using short texts with less than 200 tokens. *System* **40** (4), 554–564. 10.1016/j.system.2012.10.012 (2012).
35. Shi, R. et al. Benchmarking four large language models' performance of addressing Chinese patients' inquiries about dry eye disease: A two-phase study. *Heliyon* **10** (14), e34391. 10.1016/j.heliyon.2024.e34391 (2024). PMID: 39113991.
36. Tsai, T. H., Tsai, C. L., Hsu, J. H., Hsiao, C. H. & Chen, H. C. Bilingual evaluation of large language models for patient education in refractive surgery. *Int. J. Ophthalmol.* **19** (6), 1019–1027. 10.18240/ijo.2026.06.01 (2026). PMID: 42211204.
37. Nasra, M. et al. Can artificial intelligence improve patient educational material readability? A systematic review and narrative synthesis. *Intern. Med. J.* **55** (1), 20–34. 10.1111/imj.16607 (2025). PMID: 39720869.
38. Chandra, K. et al. Improving readability of shoulder and elbow surgery online patient education material with Chat GPT (Chat Generative Pretrained Transformer) 4. *J. Shoulder Elb. Surg.* **34** (11), e1119–e24. 10.1016/j.jse.2025.02.025 (2025). PMID: 40118438.

39. Okuhara, T., Furukawa, E., Okada, H., Yokota, R. & Kiuchi, T. Readability of written information for patients across 30 years: A systematic review of systematic reviews. Patient Educ Couns. ;135:108656. PMID: 40068244. (2025). 10.1016/j.pec.2025.108656
40. Artsi, Y. et al. Large language models in real-world clinical workflows: a systematic review of applications and implementation. Front Digit Health. ;7:1659134. PMID: 41098649. (2025). 10.3389/fdgth.2025.1659134
41. Chang, J. et al. Applications, Challenges, and Future Directions of Large Language Models in Health Care Communication: Scoping Review. *J. Med. Internet Res.* **28**, e84726. 10.2196/84726 (2026). PMID: 42361210.
42. Zhao, Y. et al. Safety guardrails in patient-facing large language model systems for chronic disease self-management: a realist review. Int J Med Inform. ;218:106513. PMID: 42224752. (2026). 10.1016/j.ijmedinf.2026.106513
43. Bedi, S. et al. Testing and Evaluation of Health Care Applications of Large Language Models: A Systematic Review. *JAMA* **333** (4), 319–328. 10.1001/jama.2024.21700 (2025). PMID: 39405325.
44. Tam, T. Y. C. et al. A framework for human evaluation of large language models in healthcare derived from literature review. NPJ Digit Med. ;7(1):258. PMID: 39333376. (2024). 10.1038/s41746-024-01258-7
45. Mokkink, L. B. et al. The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. *J. Clin. Epidemiol.* **63** (7), 737–745. 10.1016/j.jclinepi.2010.02.006 (2010). PMID: 20494804.
46. Agrawal, M., Chen, I. Y., Gulamali, F. & Joshi, S. The evaluation illusion of large language models in medicine. *NPJ Digit. Med.* **8** (1), 600. 10.1038/s41746-025-01963-x (2025). PMID: 41057566.

Figures

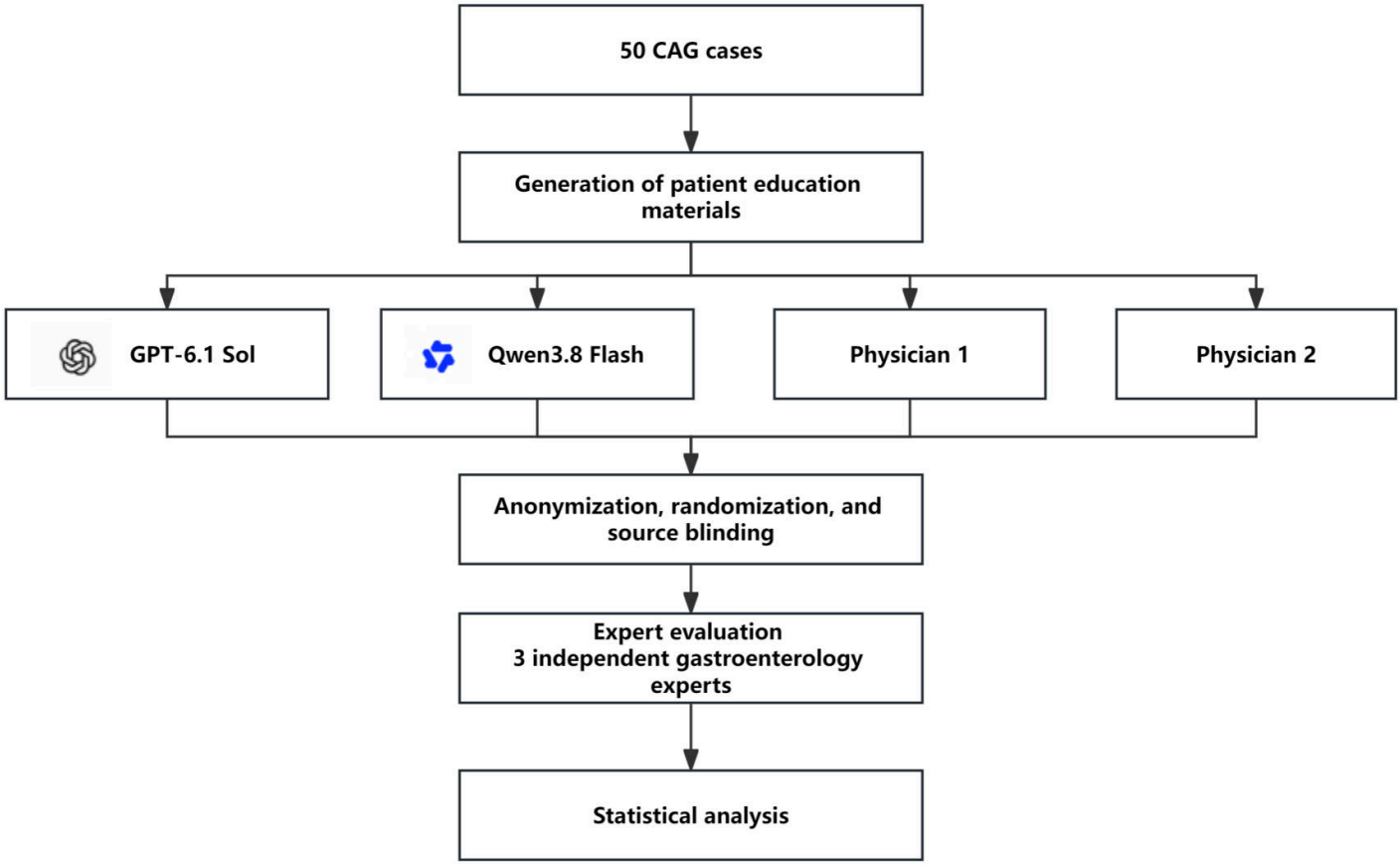


Figure 1

Study design and workflow for generation and blinded evaluation of personalized CAG patient education materials.

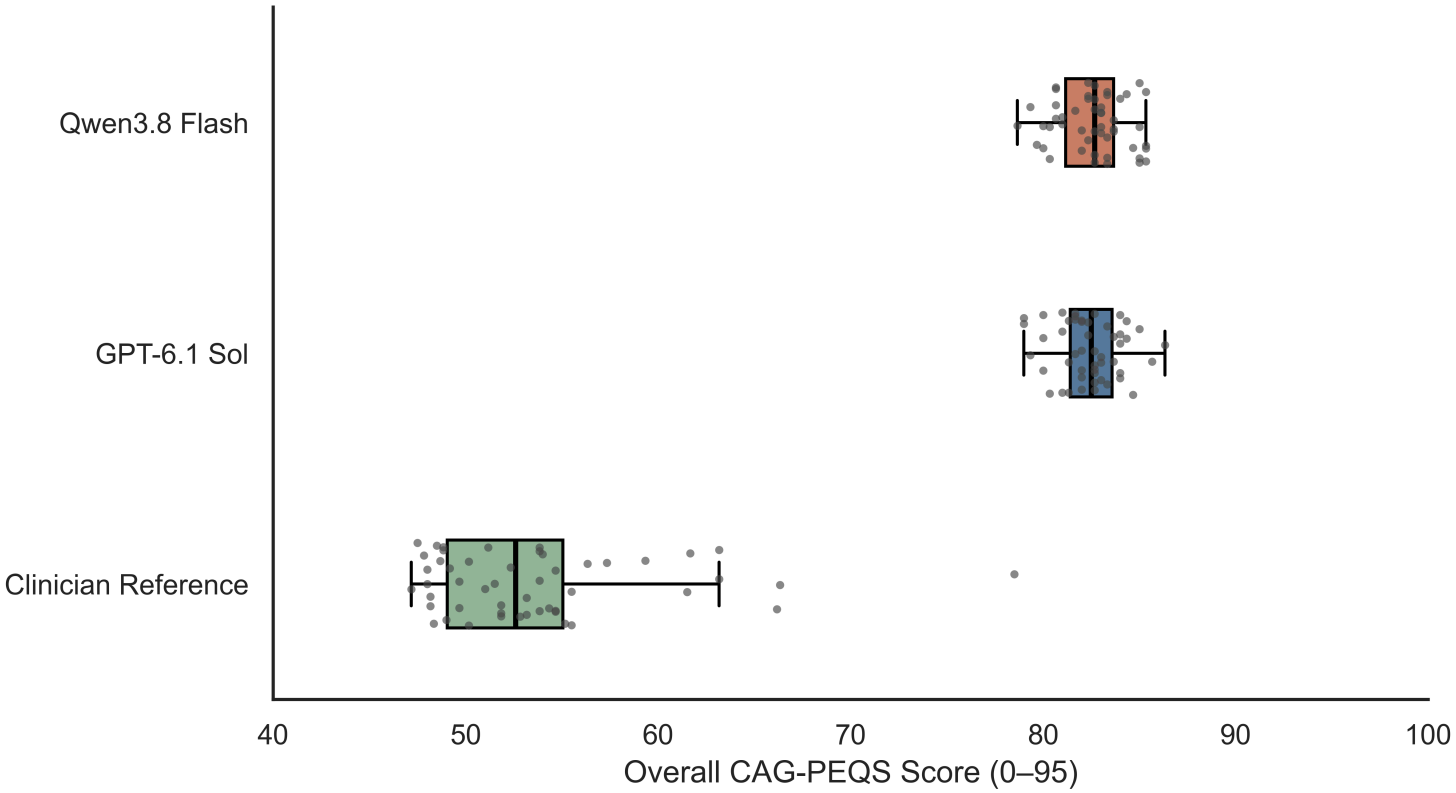


Figure 2

Distribution of overall CAG-PEQS scores across generation sources. Horizontal boxplots show case-level overall CAG-PEQS scores (0–95) for GPT-6.1 Sol, Qwen3.8 Flash, and the clinician reference across 50 matched CAG cases. Each point represents one clinical case. Boxplots indicate median and interquartile range. For each material, ratings were averaged across three independent expert reviewers. The clinician reference was calculated as the within-case mean of the two physician-generated materials.

	Medical Accuracy						Personalization						Comprehensiveness				Safety		
Qwen3.8 Flash	4.39	4.28	4.28	4.33	4.38	4.25	4.43	4.44	4.40	4.40	4.23	4.39	4.39	4.19	4.39	4.28	4.31	4.47	4.40
GPT-6.1 Sol	4.33	4.29	4.43	4.39	4.37	4.31	4.20	4.43	4.19	4.32	4.44	4.35	4.36	4.35	4.35	4.33	4.35	4.32	4.29
Clinician Reference	3.53	2.85	1.61	2.67	3.67	3.51	3.13	1.36	2.64	2.42	3.47	2.77	2.85	2.36	2.95	2.96	3.44	2.22	3.28
	A1	A2	A3	A4	A5	A6	P1	P2	P3	P4	P5	P6	C1	C2	C3	C4	S1	S2	S3

Figure 3

Item-level CAG-PEQS performance across generation sources. Each cell shows the mean item score on the 0-5 CAG-PEQS scale across 50 matched CAG cases. For each material, ratings were first averaged

across the three independent expert reviewers. The clinician reference was then calculated as the within-case mean of the two physician-generated materials. Columns are grouped into medical accuracy (A1-A6), personalization (P1-P6), comprehensiveness (C1-C4), and safety (S1-S3). A common color scale was applied to all cells, with cyan-blue indicating lower scores and coral red indicating higher scores. Full item definitions and scoring anchors are provided in Supplementary Material S2.

Supplementary Files

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

- SupplementaryMaterialS1.docx
- SupplementaryMaterialS2.docx