Original Paper

A Guideline-Concordant Chatbot Framework for Structured Colorectal Cancer Screening: Multistage Feasibility Study

Futao Wu[1*], MM; Xue Li[1*], PhD; Yingyi Zeng[1*], BMed; Yan Tang[1], MM; Zhenhua Xiao[2], MM; Siqi Yang[3], MM; Kangcheng Wu[1], BMed; Side Liu[1,4*], PhD; Aimin Li[1*], PhD

[1]Guangdong Provincial Key Laboratory of Gastroenterology, Department of Gastroenterology, Nanfang Hospital, Southern Medical University, Guangzhou, China

[2]Department of Gastroenterology, Affiliated Qingyuan Hospital, Guangzhou Medical University, Qingyuan People's Hospital, Qingyuan, China

[3]Department of Gastroenterology, Nanfang Hospital (Zengcheng Branch), Southern Medical University, Guangzhou, China

[4]Department of Gastroenterology, Zhuhai People's Hospital (Zhuhai Hospital Affiliated with Jinan University), Zhuhai, China

[*]these authors contributed equally

Corresponding Author:
Aimin Li, PhD
Guangdong Provincial Key Laboratory of Gastroenterology, Department of Gastroenterology, Nanfang Hospital, Southern Medical University
No. 1838 Guangzhou Avenue North, Baiyun District
Guangzhou, 510515
China
Phone: 86 13580317630
Email: lam0725@163.com

Abstract

Background: Colorectal cancer (CRC) screening relies on structured risk assessment and guideline-concordant communication, which remain challenging to implement consistently in real-world practice. Digital tools based on large language models (LLMs) may support such workflows, but their feasibility and safety in structured screening contexts have not been well evaluated.

Objective: This study aimed to develop and evaluate a guideline-concordant chatbot framework for structured CRC screening.

Methods: A multistage feasibility study was conducted. In phase 1, baseline performance of contemporary LLMs was assessed using 14 standardized CRC screening questions and validated expert-rated instruments. In phase 2, structured prompt versions were iteratively optimized based on screening guidelines and expert feedback and tested using simulated user scenarios. In phase 3, the optimized chatbot was evaluated in 50 screening-eligible adults to assess feasibility, safety, and guideline concordance.

Results: In phase 1, both LLMs demonstrated satisfactory performance in responding to standardized CRC screening questions, with DISCERN instrument for AI-generated content scores of 12.02 (SD 0.30) for GPT-4o and 13.36 (SD 0.26) for DeepSeek-V3, Global Quality Score scores of 3.96 (SD 0.10) for GPT-4o and 4.39 (SD 0.08) for DeepSeek-V3, Natural Language Assessment Tool for AI scores of 21.41 (SD 0.34) for GPT-4o and 22.73 (SD 0.27) for DeepSeek-V3, and Patient Education Materials Assessment Tool adapted for AI outputs scores of 0.900 (SD 0.016) for GPT-4o and 0.906 (SD 0.015) for DeepSeek-V3. In phase 2, iterative prompt optimization significantly improved all 6 expert-rated dialogue evaluation dimensions (all P values <.001). In phase 3, the optimized chatbot successfully collected complete CRC screening risk information and generated guideline-concordant screening recommendations for all participants (N=50). No unsafe or inappropriate outputs were identified.

Conclusions: This study demonstrated the feasibility, preliminary safety, and guideline-concordant performance of a structured chatbot framework for CRC screening communication under the study conditions. Beyond patient education, the proposed framework may support key components of the CRC screening workflow, including risk information collection, risk stratification, and generation of guideline-concordant screening recommendations. Further prospective studies are needed to evaluate the framework's impact on patient-centered outcomes, screening uptake, and clinical effectiveness.

(J Med Internet Res 2026;28:e93042) doi: 10.2196/93042

KEYWORDS

colorectal cancer; cancer screening; clinical decision support; digital health; artificial intelligence; AI

Introduction

Colorectal cancer (CRC) is the third most common cancer and the second leading cause of cancer death globally, with about 1.9 million new cases and 935,000 deaths annually [1]. Screening plays a key role in reducing CRC incidence and mortality by enabling early detection and removal of precancerous lesions [2,3]. However, CRC screening uptake remains suboptimal due to low awareness, misconceptions, and psychological barriers such as fear and embarrassment [4,5]. Community-based surveys indicate that real-world adherence to CRC screening ranges from approximately 13% to 55%, even in regions with established screening programs [6]. While the internet has become a major source of CRC-related information, the quality and completeness of online content remain highly variable, potentially impairing public understanding and undermining participation in screening programs [7,8].

Large language models (LLMs) are increasingly applied in health care, including patient education, clinical communication, and decision support [9-12]. Owing to factors such as open-source availability, cost-effectiveness, and suitability for local deployment, LLM-based tools have been rapidly adopted within Chinese health care systems, with reported use in more than 300 hospitals [13]. Despite this growing implementation, evidence supporting their use in CRC screening remains limited. Most existing studies have focused on answering CRC screening–related questions, providing patient education, or disseminating screening guideline information [14-16]. However, CRC screening is a highly structured clinical process that requires standardized information collection, risk assessment, and generation of guideline-concordant screening recommendations. Evidence regarding the application of LLMs within such structured screening workflows remains scarce. In addition, prior studies have raised concerns regarding the quality and reliability of AI-generated medical information [17,18]. Recent evaluations of LLMs in CRC screening have further identified inconsistencies between model-generated recommendations and current screening guidelines [16]. Given that screening eligibility and recommended screening strategies are largely determined by guideline-based criteria, ensuring alignment between LLM outputs and established recommendations is essential for clinical applicability [19].

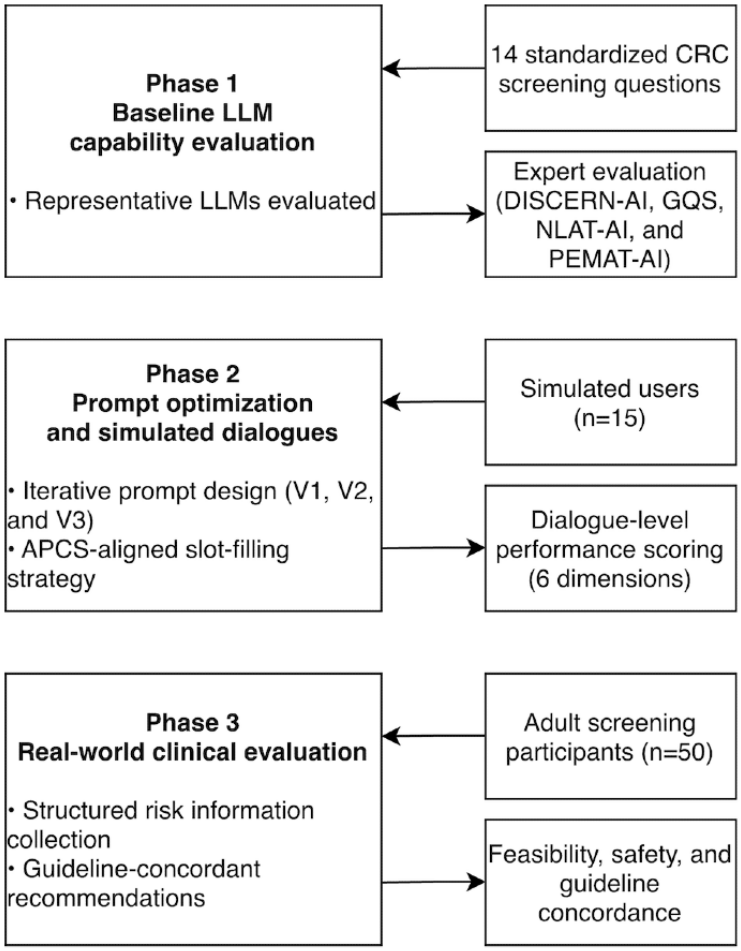
Although LLMs show promise in CRC screening, their performance in structured screening workflows requiring guideline-concordant recommendations remains insufficiently validated. Accordingly, this study aimed to develop and evaluate an LLM-based chatbot framework for CRC screening. Specifically, we aimed to (1) assess the baseline performance of contemporary LLMs in CRC screening communication; (2) optimize chatbot performance through structured prompt engineering and integration of a localized knowledge framework; and (3) evaluate the feasibility, safety, and guideline concordance of the optimized chatbot in a real-world clinical setting.

Methods

Study Design

This study was reported in accordance with the 2025 Transparency in the Reporting of Artificial Intelligence (TITAN) guideline for transparent reporting of AI use in clinical research [20]. A multiphase design was adopted to develop and evaluate an LLM-based chatbot for CRC screening decision support (Figure 1). Phase 1 assessed the baseline performance of 2 representative LLMs in responding to standardized CRC screening–related questions. Phase 2 focused on refining chatbot behavior through iterative prompt optimization and simulated dialogue testing. Phase 3 evaluated the real-world feasibility, safety, and guideline concordance of the optimized chatbot in participants undergoing clinical screening. Across all phases, model outputs and dialogue performance were independently assessed by experienced gastroenterologists using predefined and validated evaluation frameworks.

Figure 1. Overview of the multiphase framework for the development and evaluation of an large language module (LLM)–based chatbot for colorectal cancer (CRC) screening. APCS: Asia-Pacific colorectal screening; DISCERN-AI: DISCERN instrument for AI-generated content; GQS: Global Quality Score; NLAT-AI: Natural Language Assessment Tool for AI; PEMAT-AI: Patient Education Materials Assessment Tool adapted for AI outputs.



Language Model Configuration

Two representative LLMs were included in this study. GPT-4o (OpenAI), a widely used general-purpose LLM that has been extensively evaluated in health care applications, was included as a reference model for baseline comparison. DeepSeek-V3 (DeepSeek), an open-source LLM supporting local deployment, was evaluated in parallel and subsequently selected for chatbot implementation and prompt optimization.

Both models were the latest publicly available versions at the time of model evaluation and prompt development (May 2025; Table S1 in Multimedia Appendix 1). For prompt optimization, the DeepSeek-V3 model was implemented within the Cherry Studio (CherryHQ), which supported prompt design, iterative refinement, and integration of a localized knowledge base derived from established CRC screening guidelines.

Standardized Questions and Model Evaluation Procedures

A set of 14 standardized CRC screening questions was developed based on a review of the literature, expert input from gastroenterologists, Google Trends data, and key elements of China's national CRC screening program (Table S2 in Multimedia Appendix 1) [21]. The questions were designed to address major aspects of CRC screening communication, including the necessity of screening, common misunderstandings, risk-related information, and patient concerns.

To minimize potential carryover effects and contextual bias, each interaction was conducted in a newly initialized user session. Conversation history was cleared before each query, and memory-related functions were disabled when applicable to ensure a consistent and clean session state. Each standardized question was submitted as an independent single-turn query, and only the initial model-generated response was retained for analysis. All models were operated using default settings to reflect typical real-world user interactions. Representative responses are provided in Table S3 in Multimedia Appendix 1.

Simulated User Development

To assess model performance under realistic CRC screening decision-making scenarios, we constructed 15 simulated user profiles based on previously published methodologies and established CRC screening guidelines (Table S4 in Multimedia Appendix 1) [22,23]. User profiles were defined along 3 dimensions: CRC risk level (low, moderate, or high), accuracy of health-related knowledge (accurate vs inaccurate), and screening attitude (neutral vs resistant or anxious). These profiles were systematically generated to cover common combinations encountered in CRC screening practice, resulting in 12 core

profiles. In addition, several edge-case profiles incorporating incomplete or extreme information were included to evaluate model robustness. This design enabled coverage of a broad range of information completeness levels and communication challenges commonly encountered in CRC screening interactions.

Prompt Version Design (V1, V2, and V3)

Three progressively refined prompt versions (V1, V2, and V3) were developed based on prior literature, China's national CRC screening program, and feedback from preliminary testing (Table 1) [24,25]. Prompt V1 defined the chatbot's basic role as a CRC screening assistant and was used as a baseline configuration. Prompt V2 introduced a slot-filling strategy to support stepwise collection of key demographic and risk-related information for CRC screening. The collected variables were aligned with the Asia-Pacific colorectal screening (APCS) framework as applied in local screening practice, including age, sex, smoking history, BMI, and family history of CRC. Additional clinically relevant factors, such as prior screening history and alarm symptoms, were also incorporated to support risk stratification and screening decision-making. Prompt V3 further refined response generation by incorporating standardized output formatting, simplified language, and an empathetic communication style to improve clarity, interpretability, and user acceptability. Both prompt V2 and prompt V3 were integrated with a manually curated structured knowledge base derived from national CRC screening guidelines, expert consensus documents, and authoritative patient education materials to support guideline-concordant and culturally adapted responses (Table S5 in Multimedia Appendix 1).

Table 1. Iterative prompt design strategy for structured colorectal cancer screening dialogue.

Prompt version	Incremental design purpose	Incremental features
V1: role-based prompt	Establish a defined chatbot role to constrain irrelevant or generic responses.	The model is explicitly designated as a colorectal cancer screening chatbot, focusing exclusively on screening-related information and decision support.
V2: slot-filling prompt (builds on V1)	Enable structured, interactive information gathering.	Several predefined information slots are introduced, and the chatbot requests 1 missing item per turn to support stepwise risk assessment.
V3: personalized output prompt (builds on V2)	Generate personalized, understandable, and actionable advice.	Collected user data are integrated to generate personalized recommendations; output length and structure are constrained to enhance clarity, consistency, and guideline alignment.

To ensure comparability across prompt versions, a standardized user input protocol was applied. All simulated dialogues began with the same initial question ("Do I need to get screened?"), followed by predefined responses to chatbot queries. When the chatbot requested multiple pieces of information in a single turn, only the first requested item was provided according to the predefined simulation protocol. Clarification questions (eg, "Is my risk high?") were posed if no explicit risk assessment or recommendation was generated, and common follow-up concerns were introduced after recommendations. In total, 45 simulated dialogues (15 users×3 prompt versions) were generated and evaluated. Representative examples are shown in Table S6 in Multimedia Appendix 1.

Real-World Evaluation in Clinical Participants

On the basis of the optimized prompt configuration developed in the preceding phases, a CRC screening chatbot was implemented and deployed in a clinical setting (Figure S1 in Multimedia Appendix 1). Adult individuals who met the recommended criteria for CRC screening were consecutively recruited from routine screening settings.

Participants interacted with the chatbot to complete structured risk information collection and receive individualized screening recommendations (Figure S2 in Multimedia Appendix 1). All interactions were conducted in a clinical environment, with a trained physician present for observation to ensure procedural safety, without intervening in the chatbot's responses. Primary outcome measures focused on feasibility and safety. Feasibility was assessed by the successful completion of structured risk information collection and generation of individualized screening recommendations. Safety was assessed through expert review of recommendation appropriateness, guideline concordance, and the occurrence of unsafe or inappropriate outputs, consistent with prior evaluations of LLM-based tools in CRC screening and patient education [19,21]. Outputs were considered unsafe or inappropriate if they contained recommendations inconsistent with established CRC screening guidelines or were judged by expert reviewers to have the potential to adversely affect appropriate screening or follow-up. All chatbot-generated recommendations were independently reviewed by board-certified gastroenterologists using predefined evaluation criteria, and any disagreements were resolved through discussion until consensus was reached.

Evaluation Metrics and Scoring Instruments

In phase 1, model-generated responses to standardized CRC screening questions were evaluated using 4 validated assessment instruments: the Natural Language Assessment Tool for AI (NLAT-AI), the Patient Education Materials Assessment Tool adapted for AI outputs (PEMAT-AI), the modified DISCERN instrument for AI-generated content (DISCERN-AI), and the Global Quality Score (GQS) [21,26-28]. These instruments assessed key dimensions of screening-related communication, including clinical accuracy, understandability, information reliability, and overall quality. Detailed scoring criteria for each instrument are provided in Tables S7-S10 in Multimedia Appendix 1. In phase 2, chatbot-user dialogues were evaluated using a 6-dimension rubric based on a 5-point Likert scale, adapted from established evaluation frameworks [29,30]. The

evaluated dimensions included effectiveness of information gathering strategy, judgment timing appropriateness, recommendation accuracy, personalization and clarity, safety and fairness, and actionability of recommendations. All dialogue outputs were independently rated by 4 board-certified gastroenterologists under single-blind conditions. Detailed definitions and scoring criteria for each dimension are provided in Table S11 in Multimedia Appendix 1.

Statistical Analysis

Descriptive statistics were used to summarize expert ratings, baseline characteristics, and feasibility-related outcomes. Continuous variables were presented as mean (SD) or median (IQR), as appropriate, while categorical variables were summarized as counts and percentages.

Expert ratings of model-generated responses to standardized CRC screening questions were compared between models using the Wilcoxon signed-rank test. Likert-scale ratings obtained under different prompt configurations were assessed using the Kruskal-Wallis test, followed by Dunn post hoc tests with Bonferroni correction where applicable. Interrater agreement among the 4 board-certified gastroenterologists was assessed using the intraclass correlation coefficient (A,4) based on a 2-way random-effects model with absolute agreement for average measures. The real-world clinical evaluation was analyzed descriptively, with the primary aim of assessing feasibility, safety, and guideline concordance of the chatbot-assisted screening process.

All statistical analyses were performed using R software (version 4.4.1; R Foundation for Statistical Computing). A 2-sided *P* value of <.05 was considered statistically significant. Figures were generated using R software, Microsoft, and BioRender (BioRender Inc).

Ethical Considerations

Ethics approval for this study was obtained from the medical ethics committee of Nanfang Hospital, Southern Medical University (NFEC-2026-052). All participants provided written informed consent before enrollment. Access to participant data was restricted to authorized study personnel and maintained within a secure, access-controlled environment. Data were deidentified before analysis and publication to protect participant privacy and confidentiality. All study procedures were conducted in accordance with the ethical principles of the Declaration of Helsinki.

Results

Model Performance Overview

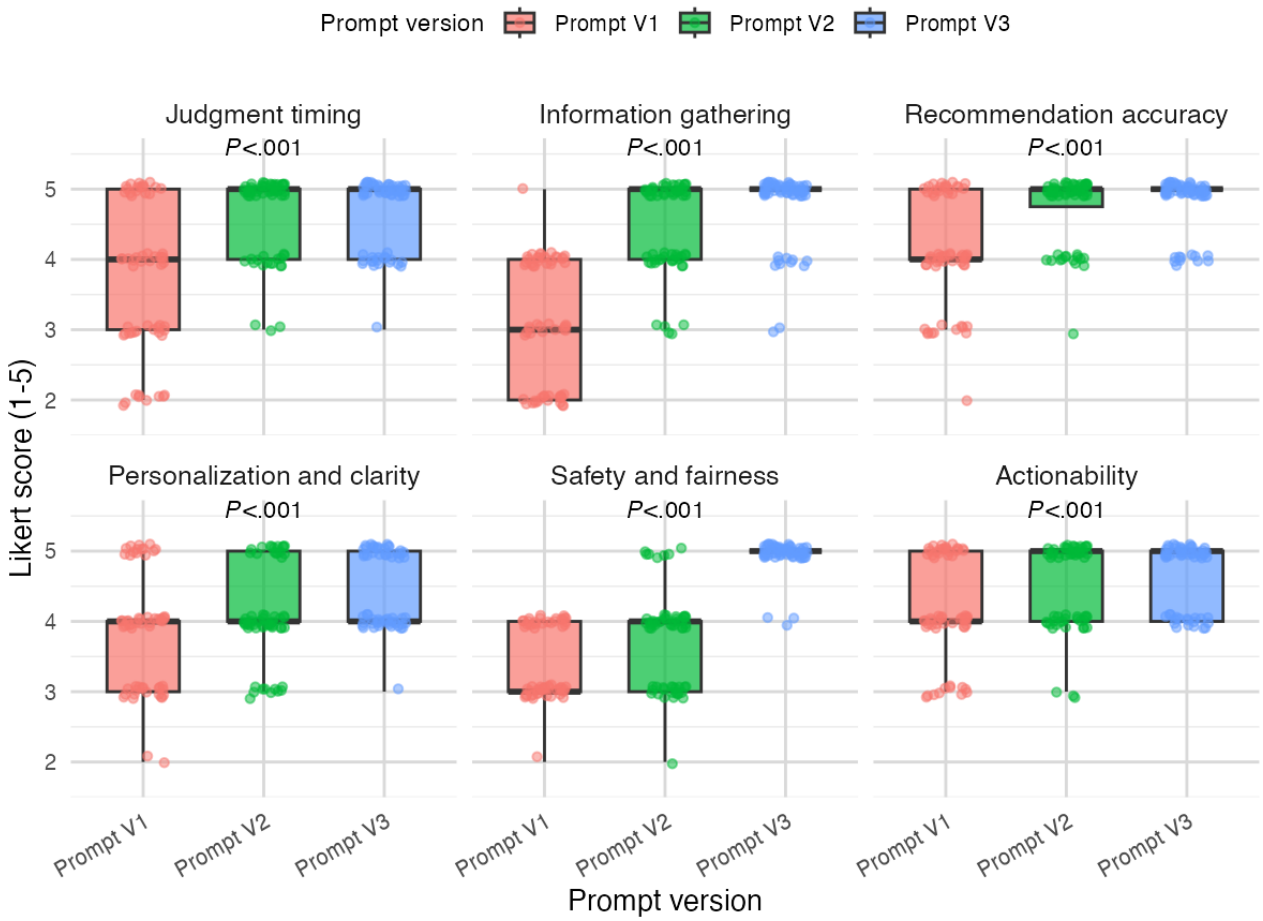
Across the 4 evaluation instruments, both LLMs demonstrated generally acceptable baseline performance in responding to standardized CRC screening questions (Figure S3 in Multimedia Appendix 1). Mean DISCERN-AI scores were 12.02 (SD 0.30) for GPT-4o and 13.36 (SD 0.26) for DeepSeek-V3, indicating generally adequate information reliability. Mean GQS values were 3.96 (SD 0.10) for GPT-4o and 4.39 (SD 0.08) for DeepSeek-V3, suggesting good overall response quality. For linguistic appropriateness, mean NLAT-AI scores were 21.41 (SD 0.34) for GPT-4o and 22.73 (SD 0.27) for DeepSeek-V3. PEMAT-AI scores were high and comparable between models (GPT-4o: mean 0.900, SD 0.016 vs DeepSeek-V3: mean 0.906, SD 0.015), reflecting similar levels of understandability and structural clarity.

Prompt Version Comparison

Interrater agreement among the 4 gastroenterologists was acceptable (intraclass correlation coefficient [A,4]=0.685, 95% CI 0.574-0.763; *P*<.001), supporting the reliability of the expert evaluation process. Boxplots illustrate the distribution of expert-rated scores across prompt versions (V1, V2, and V3) for all 6 evaluation dimensions (Figure 2). Median scores increased progressively from V1 to V3 across all dimensions, accompanied by reduced score dispersion in later prompt versions, particularly for information gathering, recommendation accuracy, and safety and fairness. Overall differences among prompt versions were statistically significant across all 6 dimensions (Kruskal-Wallis test, all *P* values <.001). Post hoc analyses using Dunn test with Bonferroni correction showed that prompt V3 scored higher than prompt V1 across all dimensions (adjusted *P*<.05). Compared with prompt V2, prompt V3 demonstrated statistically significant improvements only in personalization and clarity (adjusted *P*=.02) and safety and fairness (adjusted *P*<.001), whereas differences in the remaining dimensions did not reach statistical significance (adjusted *P*>.05).

Radar plots summarize the mean scores of each prompt version across the 6 evaluation dimensions (Figure S4 in Multimedia Appendix 1). From a descriptive perspective, mean scores increased consistently from V1 to V3 across all dimensions, with no dimension exhibiting a decrease in mean performance.

Figure 2. Expert-rated dialogue performance across iterative prompt versions for colorectal cancer screening.



Real-World Clinical Evaluation

A total of 50 clinical participants were enrolled to assess the real-world feasibility of the optimized CRC screening chatbot (Table 2). Among them, 46% (n=23) were male and 76% (n=38) were aged ≥50 years. Complete CRC screening risk information was collected for all the participants through chatbot-assisted interactions, and all generated screening recommendations were guideline concordant.

APCS-based risk stratification was concordant with the chatbot's final risk assessment in 92% (46/50) of cases. Expert review confirmed that all chatbot-generated screening recommendations were consistent with current screening guidelines, and no inappropriate or unsafe outputs were identified during the study period.

Table 2. Baseline characteristics and feasibility outcomes of real-world chatbot-assisted colorectal cancer (CRC) screening.

	Values (N=50)
Baseline characteristics	
Age (years), mean (SD)	56.8 (8.91)
Age ≥50 years, n (%)	38 (76)
Sex (male), n (%)	23 (46)
BMI (kg/m²), mean (SD)	22.9 (3.85)
Prior CRC screening, n (%)	15 (30)
Family history of CRC, n (%)	3 (6)
Feasibility outcomes of chatbot-assisted CRC screening, n (%)	
Complete risk information collected	50 (100)
Concordance between APCS[a]-based risk stratification and final risk assessment	46 (92)
Guideline-concordant screening recommendation	50 (100)
Unsafe or inappropriate outputs	0 (0)

[a]APCS: Asia-Pacific colorectal screening.

Discussion

Principal Findings

To our knowledge, this is among the first studies to develop and clinically evaluate a chatbot framework for structured CRC screening. Unlike previous studies that primarily focused on patient education or responses to screening-related questions, our framework was designed to support the entire screening workflow, including risk information collection, risk stratification, and generation of guideline-concordant screening recommendations. To achieve this, we combined iterative prompt optimization with a localized guideline-based knowledge base and evaluated the resulting system in a real-world clinical setting. Together, these findings provide preliminary evidence supporting the feasibility of using LLM-based systems as workflow-oriented screening assistants for guideline-driven CRC screening under the study conditions.

Previous studies have demonstrated that LLMs can provide accurate, complete, and understandable responses to CRC screening–related questions, highlighting their potential value in patient education and health communication [19,31]. In contrast, this study evaluated the potential of LLMs to support structured CRC screening workflows rather than question answering alone. The standardized question set used in our study was designed to cover key components of this workflow, including screening eligibility assessment, risk information collection, risk stratification, screening modality selection, bowel preparation, result interpretation, and follow-up management. Accordingly, our evaluation extended beyond information provision and assessed the potential of LLMs to support structured screening communication and workflow-oriented decision processes. Within this framework, both GPT-4o and DeepSeek-V3 demonstrated satisfactory performance across DISCERN-AI, GQS, NLAT-AI, and PEMAT-AI, suggesting that contemporary LLMs possess not only educational value but also the foundational capabilities required to support structured CRC screening workflows (Figure S5 in Multimedia Appendix 1).

DeepSeek-V3 was selected as the backbone model for CRC screening support because of its open-source availability and suitability for local deployment [32]. These features are particularly relevant for public health applications requiring data privacy, regulatory compliance, and language adaptation [33,34]. Nevertheless, expert reviewers noted occasional use of technical terminology without sufficient explanation, as well as variability in response structure, particularly in follow-up management scenarios. Similar limitations have been reported in previous evaluations of LLMs for CRC screening, where model-generated recommendations did not always align with current screening guidelines and occasionally relied on outdated screening criteria or oversimplified recommendations for high-risk populations [16]. Prior studies have shown that integrating clinical guideline knowledge into LLMs can improve their ability to generate accurate and guideline-concordant responses in medical tasks [24]. Collectively, these observations suggest that translating the general capabilities of LLMs into structured screening support remains challenging and highlight the value of incorporating specialized clinical knowledge into screening-oriented chatbot systems. Our findings further suggest that performance in highly protocol-driven tasks, such as cancer screening, depends not only on the underlying language model but also on how screening workflows and clinical knowledge are structured and integrated into the dialogue system. The progressive improvements observed during iterative prompt refinement, particularly in information completeness and safety, suggest that careful dialogue design can substantially enhance the quality and guideline concordance of screening-related communication. These findings suggest that combining foundation models with domain-specific knowledge and workflow-oriented design may facilitate the development of screening support systems, although further validation is required before clinical implementation and clinical effectiveness can be established. When implemented within a

locally deployable, open-source framework, such approaches may facilitate the development of culturally adaptable AI-assisted tools for CRC screening that can be aligned with local screening guidelines, languages, and health care systems. Such characteristics may be particularly relevant for health care settings with stringent requirements for data governance, regulatory compliance, and guideline concordance [35].

In real-world clinical evaluation, the chatbot-generated risk assessment was concordant with APCS-based stratification in most cases, with discrepancies observed in only a small subset of participants. These discrepancies were not attributable to system errors or human intervention. Instead, they reflected the model's incorporation of clinically relevant contextual information not explicitly captured by the APCS scoring system. Specifically, 2 discordant cases involved participants reporting alarm symptoms, and 1 case involved a participant with a previously positive fecal occult blood test that had not been followed up. In these cases, the chatbot assigned a higher risk assessment than APCS-based stratification and recommended diagnostic colonoscopy. In contrast, 1 discordant case involved a participant classified as high risk according to the APCS who had undergone a normal colonoscopy 1 year earlier. In this case, the chatbot assigned a lower risk assessment and generated a recommendation based on the guideline-recommended surveillance interval. Together, these cases highlight the potential value of incorporating additional clinically relevant information beyond APCS-based stratification for individualized risk assessment. Importantly, all resulting screening recommendations remained consistent with current CRC screening guidelines. This finding highlights an important distinction between population-level risk stratification tools and individualized screening decision support. While APCS is designed to categorize baseline risk, screening recommendations in clinical practice often require integration of additional contextual information beyond the risk score alone. In this study, the chatbot incorporated this information through dialogue-based interactions rather than relying solely on risk score calculation. Although expert review confirmed that all chatbot-generated recommendations remained guideline concordant, the clinical value of incorporating additional contextual information beyond APCS-based risk stratification remains uncertain. Future prospective studies with longitudinal follow-up will be needed to determine whether such recommendation adjustments improve screening outcomes and provide incremental benefit beyond conventional risk stratification approaches. In addition, sustaining the performance of guideline-based chatbot systems in real-world practice will require ongoing alignment between the underlying knowledge base and evolving CRC screening recommendations. Future implementations may incorporate structured expert review and revision of the knowledge base following major guideline updates, along with targeted scenario testing to ensure the continued accuracy and guideline concordance of chatbot-generated recommendations.

This study has several strengths. First, we adopted a stepwise, multistage study design, in which the CRC screening chatbot evaluated in the final phase was systematically developed and optimized based on findings from earlier phases, including baseline model assessment and iterative prompt refinement. Second, by integrating CRC screening guideline knowledge with structured prompt engineering, the chatbot was specifically designed to support a highly protocol-driven screening workflow rather than general health question answering. Third, chatbot outputs were independently evaluated in a single-blind manner by multiple experienced gastroenterologists using validated assessment instruments, supporting the reliability and clinical relevance of the findings. Finally, the use of an open-source and locally deployable LLM may support future adaptation of this framework to other screening and preventive care applications that require guideline-concordant recommendations.

This study has several limitations. First, although real clinical participants were included, the sample size was modest, and the study was designed to assess feasibility and safety rather than comparative clinical effectiveness. Accordingly, direct comparisons with standard clinician counseling, existing screening tools, or usual care were not performed. Second, although chatbot outputs were independently reviewed by experienced gastroenterologists using predefined evaluation criteria, assessment of guideline concordance and clinical appropriateness inevitably involved some degree of subjectivity. Third, patient-centered outcomes, including understanding of screening recommendations, decision confidence, screening uptake, patient behavior, and long-term clinical end points, were not examined. In addition, safety was assessed through expert review of chatbot-generated outputs rather than downstream patient outcomes. Therefore, this study cannot determine whether chatbot use influences delayed care, screening uptake, or other real-world clinical consequences. Fourth, chatbot interactions were conducted in the presence of a trained physician for observation. Although no intervention occurred during chatbot use, this supervised setting may differ from unsupervised real-world deployment and may limit the generalizability of the findings. Finally, as the system was developed and evaluated in a Chinese-language screening context, further validation is required to establish generalizability across other languages and health care systems. Despite these limitations, this work represents an initial step in a sequential research program. Future prospective randomized controlled trials are planned to evaluate the impact of the optimized CRC screening chatbot on patient-centered outcomes, screening uptake, and clinical effectiveness.

Conclusions

This study developed and evaluated an LLM-based chatbot for CRC screening. By integrating structured prompt engineering with a localized guideline-derived knowledge base, the system demonstrated feasibility, preliminary safety, and guideline-concordant performance under the study conditions. The proposed framework provides a foundation for future research on AI-assisted support in structured screening workflows and may inform the development of similar systems for other clinical tasks requiring standardized information collection and guideline-based recommendations.

Acknowledgments

The authors would like to thank all the participants for their time and valuable contributions to this study. In addition, the authors acknowledge Guangzhou Golden Image Information Technology Co., Ltd (Goldenimg) for providing technical support during the development and testing of the chatbot prototype used in this study. The company assisted with system implementation and platform integration only and had no role in the study design, data analysis, interpretation of the results, or preparation of the manuscript. The generative artificial intelligence tool ChatGPT (OpenAI [36]) was used solely as a grammar and language editing tool to refine the clarity and flow of the original manuscript. It played no role in the study conception, data analysis, or substantive drafting of the research.

SL is a co-corresponding author of this paper and can be reached at liuside2011@163.com.

Funding

This work was supported by the National Natural Science Foundation of China (82373165); the Guangdong Provincial Key Laboratory of Gastroenterology (2017B030314037); the Sanming Project of Medicine in Shenzhen (SZSM202105017); the Key Medical Discipline Program of Longgang District, Shenzhen; and the Guangzhou Clinical High-level, Major, and Characteristic Medical Technology Project (2026P-ZD022). The funder had no involvement in the study design, data collection, analysis, interpretation, or the writing of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Supplementary materials, including technical specifications of the large language models, standardized evaluation questions, prompt versions, simulated user profiles, representative conversations, evaluation instruments, supplementary figures and tables, and additional study results.
[DOCX File , 398 KB-Multimedia Appendix 1]

References

1. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. May 2021;71(3):209-249. [FREE Full text] [doi: 10.3322/caac.21660] [Medline: 33538338]
2. US Preventive Services Task Force, Davidson KW, Barry MJ, Mangione CM, Cabana M, Caughey AB, et al. Screening for colorectal cancer: US Preventive Services Task Force recommendation statement. JAMA. May 18, 2021;325(19):1965-1977. [doi: 10.1001/jama.2021.6238] [Medline: 34003218]
3. Zauber AG, Winawer SJ, O'Brien MJ, Lansdorp-Vogelaar I, van Ballegooijen M, Hankey BF, et al. Colonoscopic polypectomy and long-term prevention of colorectal-cancer deaths. N Engl J Med. Feb 23, 2012;366(8):687-696. [FREE Full text] [doi: 10.1056/NEJMoa1100370] [Medline: 22356322]
4. Amlani B, Radaelli F, Bhandari P. A survey on colonoscopy shows poor understanding of its protective value and widespread misconceptions across Europe. PLoS One. May 21, 2020;15(5):e0233490. [FREE Full text] [doi: 10.1371/journal.pone.0233490] [Medline: 32437402]
5. Cossu G, Saba L, Minerba L, Mascalchi M. Colorectal cancer screening: the role of psychological, social and background factors in decision-making process. Clin Pract Epidemiol Ment Health. Mar 21, 2018;14:63-69. [FREE Full text] [doi: 10.2174/1745017901814010063] [Medline: 29643929]
6. Kew GS, Koh CJ. Strategies to improve persistent adherence in colorectal cancer screening. Gut Liver. Sep 15, 2020;14(5):546-552. [FREE Full text] [doi: 10.5009/gnl19306] [Medline: 31822055]
7. Ruco A, Dossa F, Tinmouth J, Llovet D, Jacobson J, Kishibe T, et al. Social media and mHealth technology for cancer screening: systematic review and meta-analysis. J Med Internet Res. Jul 30, 2021;23(7):e26759. [FREE Full text] [doi: 10.2196/26759] [Medline: 34328423]
8. Wang L, Gusnowski EM, Ingledew PA. Digesting the contents: an analysis of online colorectal cancer education websites. J Cancer Educ. Apr 2022;37(2):263-273. [doi: 10.1007/s13187-020-01864-5] [Medline: 32902788]
9. Bedi S, Liu Y, Orr-Ewing L, Dash D, Koyejo S, Callahan A, et al. Testing and evaluation of health care applications of large language models: a systematic review. JAMA. Jan 28, 2025;333(4):319-328. [doi: 10.1001/jama.2024.21700] [Medline: 39405325]
10. Pan A, Musheyev D, Bockelman D, Loeb S, Kabarriti AE. Assessment of artificial intelligence chatbot responses to top searched queries about cancer. JAMA Oncol. Oct 01, 2023;9(10):1437-1440. [doi: 10.1001/jamaoncol.2023.2947] [Medline: 37615960]

11. Zhou Z, Qin P, Cheng X, Shao M, Ren Z, Zhao Y, et al. ChatGPT in oncology diagnosis and treatment: applications, legal and ethical challenges. Curr Oncol Rep. Apr 2025;27(4):336-354. [doi: 10.1007/s11912-025-01649-3] [Medline: 39998782]
12. Gan W, Ouyang J, Li H, Xue Z, Zhang Y, Dong Q, et al. Integrating ChatGPT in orthopedic education for medical undergraduates: randomized controlled trial. J Med Internet Res. Aug 20, 2024;26:e57037. [FREE Full text] [doi: 10.2196/57037] [Medline: 39163598]
13. Zeng D, Qin Y, Sheng B, Wong TY. DeepSeek's "low-cost" adoption across China's hospital systems: too fast, too soon? JAMA. Jun 03, 2025;333(21):1866-1869. [doi: 10.1001/jama.2025.6571] [Medline: 40293869]
14. M M, A P, S G, T V, Lhs L, S B, et al. World Endoscopy Organization (WEO) Emerging Stars Program. Performance of large language models in addressing patient queries on colorectal cancer screening in different languages: an international study across 28 countries. Dig Liver Dis. Feb 2026;58(2):250-257. [FREE Full text] [doi: 10.1016/j.dld.2025.11.026] [Medline: 41436291]
15. Maida M, Mori Y, Fuccio L, Sferrazza S, Vitello A, Facciorusso A, et al. Exploring ChatGPT effectiveness in addressing direct patient queries on colorectal cancer screening. Endosc Int Open. May 12, 2025;13:a25689416. [FREE Full text] [doi: 10.1055/a-2568-9416] [Medline: 40376022]
16. Zhang Z, Zhang ZC, Zhang SP, Luan WY, Han S, Xu ZX, et al. Comparative analysis of artificial intelligence tools for the dissemination of colorectal cancer screening guidelines: a novel perspective on early screening education. Int J Surg. Nov 01, 2025;111(11):8616-8620. [doi: 10.1097/JS9.0000000000002951] [Medline: 40607944]
17. Peng Y, Malin BA, Rousseau JF, Wang Y, Xu Z, Xu X, et al. From GPT to DeepSeek: significant gaps remain in realizing AI in healthcare. J Biomed Inform. Mar 2025;163:104791. [FREE Full text] [doi: 10.1016/j.jbi.2025.104791] [Medline: 39938624]
18. Zhou M, Pan Y, Zhang Y, Song X, Zhou Y. Evaluating AI-generated patient education materials for spinal surgeries: comparative analysis of readability and DISCERN quality across ChatGPT and DeepSeek models. Int J Med Inform. Jun 2025;198:105871. [FREE Full text] [doi: 10.1016/j.ijmedinf.2025.105871] [Medline: 40107040]
19. Atarere J, Naqvi H, Haas C, Adewunmi C, Bandaru S, Allamneni R, et al. Applicability of online chat-based artificial intelligence models to colorectal cancer screening. Dig Dis Sci. Mar 2024;69(3):791-797. [doi: 10.1007/s10620-024-08274-3] [Medline: 38267726]
20. Agha RA, Mathew G, Rashid R, Kerwan A, Al-Jabir A, Sohrabi C, et al. Transparency In The reporting of Artificial INtelligence – the TITAN guideline. Premier J Sci. 2025;10:100082. [doi: 10.70389/PJS.100082]
21. Siu AH, Gibson DP, Chiu C, Kwok A, Irwin M, Christie A, et al. ChatGPT as a patient education tool in colorectal cancer-an in-depth assessment of efficacy, quality and readability. Colorectal Dis. Jan 2025;27(1):e17267. [doi: 10.1111/codi.17267] [Medline: 39690137]
22. Kaygisiz Ö, Teke MT. Can DeepSeek and ChatGPT be used in the diagnosis of oral pathologies? BMC Oral Health. Apr 25, 2025;25(1):638. [FREE Full text] [doi: 10.1186/s12903-025-06034-x] [Medline: 40281436]
23. Hswen Y, Nguyen T. Assessing variability in ChatGPT responses: a case study on simulating online user inputs. J Assoc Nurses AIDS Care. Apr 29, 2025. (forthcoming). [doi: 10.1097/JNC.0000000000000545] [Medline: 40298212]
24. Lim DY, Tan YB, Koh JT, Tung JY, Sng GG, Tan DM, et al. ChatGPT on guidelines: providing contextual knowledge to GPT allows it to provide advice on appropriate colonoscopy intervals. J Gastroenterol Hepatol. Jan 2024;39(1):81-106. [doi: 10.1111/jgh.16375] [Medline: 37855067]
25. Holderried F, Stegemann-Philipps C, Herschbach L, Moldt JA, Nevins A, Griewatz J, et al. A generative pretrained transformer (GPT)-powered chatbot as a simulated patient to practice history taking: prospective, mixed methods study. JMIR Med Educ. Jan 16, 2024;10:e53961. [FREE Full text] [doi: 10.2196/53961] [Medline: 38227363]
26. Charnock D, Shepperd S, Needham G, Gann R. DISCERN: an instrument for judging the quality of written consumer health information on treatment choices. J Epidemiol Community Health. Feb 1999;53(2):105-111. [FREE Full text] [doi: 10.1136/jech.53.2.105] [Medline: 10396471]
27. Dastani M, Mardaneh J, Rostamian M. Large language models' capabilities in responding to tuberculosis medical questions: testing ChatGPT, Gemini, and Copilot. Sci Rep. May 23, 2025;15(1):18004. [FREE Full text] [doi: 10.1038/s41598-025-03074-9] [Medline: 40410343]
28. Gibson D, Jackson S, Shanmugasundaram R, Seth I, Siu A, Ahmadi N, et al. Evaluating the efficacy of ChatGPT as a patient education tool in prostate cancer: multimetric assessment. J Med Internet Res. Aug 14, 2024;26:e55939. [FREE Full text] [doi: 10.2196/55939] [Medline: 39141904]
29. Fink MA, Bischoff A, Fink CA, Moll M, Kroschke J, Dulz L, et al. Potential of ChatGPT and GPT-4 for data mining of free-text CT reports on lung cancer. Radiology. Sep 2023;308(3):e231362. [doi: 10.1148/radiol.231362] [Medline: 37724963]
30. Goodman RS, Patrinely JR, Stone CAJ, Zimmerman E, Donald RR, Chang SS, et al. Accuracy and reliability of chatbot responses to physician questions. JAMA Netw Open. Oct 02, 2023;6(10):e2336483. [FREE Full text] [doi: 10.1001/jamanetworkopen.2023.36483] [Medline: 37782499]
31. Maida M, Ramai D, Mori Y, Dinis-Ribeiro M, Facciorusso A, Hassan C. The role of generative language systems in increasing patient awareness of colon cancer screening. Endoscopy. Mar 2025;57(3):262-268. [FREE Full text] [doi: 10.1055/a-2388-6084] [Medline: 39142348]

32. Kayaalp ME, Prill R, Sezgin EA, Cong T, Królikowska A, Hirschmann MT. DeepSeek versus ChatGPT: multimodal artificial intelligence revolutionizing scientific discovery. From language editing to autonomous content generation-redefining innovation in research and practice. Knee Surg Sports Traumatol Arthrosc. May 2025;33(5):1553-1556. [doi: 10.1002/ksa.12628] [Medline: 39936363]
33. Sandmann S, Hegselmann S, Fujarski M, Bickmann L, Wild B, Eils R, et al. Benchmark evaluation of DeepSeek large language models in clinical decision-making. Nat Med. Aug 2025;31(8):2546-2549. [doi: 10.1038/s41591-025-03727-2] [Medline: 40267970]
34. Wang C, Liu S, Yang H, Guo J, Wu Y, Liu J. Ethical considerations of using ChatGPT in health care. J Med Internet Res. Aug 11, 2023;25:e48009. [FREE Full text] [doi: 10.2196/48009] [Medline: 37566454]
35. Temsah A, Alhasan K, Altamimi I, Jamal A, Al-Eyadhy A, Malki KH, et al. DeepSeek in healthcare: revealing opportunities and steering challenges of a new open-source artificial intelligence frontier. Cureus. Feb 18, 2025;17(2):e79221. [doi: 10.7759/cureus.79221] [Medline: 39974299]
36. ChatGPT. URL: https://chatgpt.com/ [accessed 2026-07-06]

Abbreviations

APCS: Asia-Pacific colorectal screening
CRC: colorectal cancer
DISCERN-AI: DISCERN instrument for AI-generated content
GQS: Global Quality Score
LLM: large language model
NLAT-AI: Natural Language Assessment Tool for AI
PEMAT-AI: Patient Education Materials Assessment Tool adapted for AI outputs
TITAN: Transparency in the Reporting of Artificial Intelligence

Edited by M Balcarras; submitted 06.Feb.2026; peer-reviewed by F Liu, M-H Tsai; comments to author 15.Jun.2026; accepted 25.Jun.2026; published 16.Jul.2026

Please cite as:
Wu F, Li X, Zeng Y, Tang Y, Xiao Z, Yang S, Wu K, Liu S, Li A
A Guideline-Concordant Chatbot Framework for Structured Colorectal Cancer Screening: Multistage Feasibility Study
J Med Internet Res 2026;28:e93042
URL: https://www.jmir.org/2026/1/e93042
doi: 10.2196/93042
PMID: