

Assessing ChatGPT4 with and without retrieval-augmented generation in anticoagulation management for gastrointestinal procedures

Sheza Malik^a, Himal Kharel^a, Dushyant S Dahiya^b, Hassam Ali^c, Hanna Blaney^d, Achintya Singh^e, Jahnvi Dhar^f, Abhilash Perisetti^g, Antonio Facciorusso^h, Saurabh Chandanⁱ, Babu P. Mohan^j

Rochester General Hospital, NY, USA; University of Kansas School of Medicine, Kansas, USA; East Carolina University, NC, USA; New York University Grossman School of Medicine, NC, USA; Metro Health, OH, USA; Postgraduate Institute of Medical Education and Research, Chandigarh, India; Kansas City VA, USA; University of Foggia, Italy; Creighton University Medical Center, USA; Orlando Gastroenterology, FL, USA

Abstract

Background In view of the growing complexity of managing anticoagulation for patients undergoing gastrointestinal (GI) procedures, this study evaluated ChatGPT-4's ability to provide accurate medical guidance, comparing it with its prior artificial intelligence (AI) models (ChatGPT-3.5) and the retrieval-augmented generation (RAG)-supported model (ChatGPT4-RAG).

Methods Thirty-six anticoagulation-related questions, based on professional guidelines, were answered by ChatGPT-4. Nine gastroenterologists assessed these responses for accuracy and relevance. ChatGPT-4's performance was also compared to that of ChatGPT-3.5 and ChatGPT4-RAG. Additionally, a survey was conducted to understand gastroenterologists' perceptions of ChatGPT-4.

Results ChatGPT-4's responses showed significantly better accuracy and coherence compared to ChatGPT-3.5, with 30.5% of responses fully accurate and 47.2% generally accurate. ChatGPT4-RAG demonstrated a higher ability to integrate current information, achieving 75% full accuracy. Notably, for diagnostic and therapeutic esophagogastroduodenoscopy, 51.8% of responses were fully accurate; for endoscopic retrograde cholangiopancreatography with and without stent placement, 42.8% were fully accurate; and for diagnostic and therapeutic colonoscopy, 50% were fully accurate.

Conclusions ChatGPT4-RAG significantly advances anticoagulation management in endoscopic procedures, offering reliable and precise medical guidance. However, medicolegal considerations mean that a 75% full accuracy rate remains inadequate for independent clinical decision-making. AI may be more appropriately utilized to support and confirm clinicians' decisions, rather than replace them. Further evaluation is essential to maintain patient confidentiality and the integrity of the physician-patient relationship.

Keywords Anticoagulation management, gastrointestinal procedures, accuracy, ChatGPT4-RAG, endoscopic procedures

Ann Gastroenterol 2024; 37 (XX): 1-13

Conflict of Interest: None

Correspondence to: Sheza Malik, Rochester General Hospital, Portland Avenue, Rochester, NYC, USA 14621, e-mail: Sheza.malik683@gmail.com

Received 1 May 2024; accepted 27 June 2024; published online 19 August 2024

DOI: <https://doi.org/10.20524/aog.2024.0907>

This is an open access journal, and articles are distributed under the terms of the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 License, which allows others to remix, tweak, and build upon the work non-commercially, as long as appropriate credit is given and the new creations are licensed under the identical terms

Introduction

The increasing demand for gastrointestinal (GI) procedures in the context of a growing population on anticoagulation therapy presents a complex challenge for healthcare systems worldwide [1,2]. As the prevalence of conditions requiring anticoagulation therapy rises, so does the complexity of managing these patients before undergoing GI procedures, leading to inquiries and concerns from primary care physicians and patients alike [3]. These inquiries often encompass anticoagulation therapy's safety, timing and management, highlighting the need for clear, accessible and accurate medical guidance [4].

In the face of these challenges, the potential of artificial intelligence (AI), particularly ChatGPT, emerges as a promising solution to support GI physicians by providing immediate, reliable answers to common anticoagulation-related questions [5]. Integrating AI into the clinical setting could alleviate the burden on healthcare professionals, enhance patient education, and streamline the pre-procedure preparation process. By leveraging the advanced capabilities of ChatGPT, GI specialists can access a tool that aids in decision-making, ensuring that patients receive timely and appropriate care [6].

Large language models (LLMs) like ChatGPT have advanced significantly in recent years, yet they encounter challenges such as hallucinations, where they generate misleading or incorrect information. This issue stems from the inherent limitations of their training processes, rather than intentional misinformation. Researchers are improving data quality, modifying training methods, and utilizing real-world fact checks to tackle this. Among these strategies, retrieval-augmented generation (RAG) stands out by merging LLMs' generative abilities with an information retrieval process, allowing models to consult a factual database during generation. This technique grounds the models' outputs in real-world data, significantly reducing the tendency for hallucinated content and enhancing the reliability of their responses [7,8]. Fig. 1 explains the functioning of RAG. The latest version of ChatGPT allows users to submit their documents for the model to reference, improving the reliability of the generated information (ChatGPT4-RAG).

This study aimed to evaluate the effectiveness of ChatGPT4, with and without RAG, in responding to anticoagulation management questions pertinent to endoscopic procedures, assessing its accuracy, coherence and medical relevance compared to existing AI models and versions. Through expert evaluations and a survey of gastroenterologists' perceptions, the research sought to establish ChatGPT's role in augmenting care delivery in gastroenterology.

Materials and methods

Data source

We carefully formulated questions about anticoagulation management before endoscopic procedures, basing them on

^aInternal Medicine, Rochester General Hospital, NY, USA (Sheza Malik, Himal Kharel); ^bGastroenterology, Hepatology, University of Kansas School of Medicine, Kansas, USA (Dushyant S. Dahiya); ^cGastroenterology, Hepatology, East Carolina University, NC, USA (Hassam Ali); ^dGastroenterology, Hepatology, New York University Grossman School of Medicine, NC, USA (Hanna Blaney); ^eGastroenterology, Hepatology, Metro Health, OH, USA (Achintya Singh); ^fGastroenterology, Hepatology, Postgraduate Institute of Medical Education and Research, Chandigarh, India (Jahnvi Dhar); ^gGastroenterology, Hepatology, Kansas City VA, USA (Abhilash Periseti); ^hGastroenterology, Hepatology, University of Foggia, Italy (Antonio Facciorusso); ⁱGastroenterology, Hepatology, Creighton University Medical Center, USA (Saurabh Chandan); ^jGastroenterology, Hepatology, Orlando Gastroenterology, FL, USA (Babu P. Mohan)

a professional society's (American Society of Gastrointestinal Endoscopy) guidelines [1]. The criteria for exclusion consisted of questions that conveyed similar meanings, questions with ambiguous meanings (such as inquiries about how endoscopic procedures impact the body), queries that could differ from individual to individual (such as the likelihood of a person's condition worsening after the procedure), and questions not related to the medical aspects of the procedures. A total of 36 questions were selected for common endoscopy procedures: 6 questions each for diagnostic and therapeutic esophagogastroduodenoscopy (EGD); endoscopic retrograde cholangiopancreatography (ERCP), with and without stent placement; endoscopic ultrasound (EUS) with fine-needle aspiration (FNA); diagnostic and therapeutic colonoscopy; percutaneous endoscopic gastrostomy (PEG); and enteral stent deployment).

Response generation

In ChatGPT, answers are generated through advanced natural language processing techniques, leveraging an internet-scale training dataset and incorporating reinforcement learning from human feedback to refine its output, ensuring high-quality and contextually accurate responses. Queries (in English) were entered by an author (SM) into the most recent iteration of ChatGPT 4, using the platform's "New Chat" functionality, ensuring each prompt was treated as a distinct and isolated input [7]. The responses generated were meticulously documented in Google Docs (Table 1) and sent to the evaluators (gastroenterologists) for grading. We prioritized ChatGPT over other models such as BLOOM (BigScience, various institutions), LaMBDA/Bard (Google, Mountain View, Calif), and LLaMA (Meta AI, Menlo Park, Calif), in view of its established reputation, extensive training data, seamless integration and widespread recognition [9,10].

Grading of questions

Ten gastroenterologists (5 fellows and 5 consultant gastroenterologists) from the USA, Italy and India were tasked with evaluating the medical accuracy of each response generated in the study. All participants demonstrated fluency in the English language. Their assessments were based on predefined keys that categorized responses into 3 levels of accuracy: "Fully accurate", indicating complete alignment with established medical facts and knowledge; "Generally accurate", where responses, despite potentially minor inaccuracies or omissions, were largely correct; and "Predominantly or completely inaccurate", signifying significant deviations from accepted medical understanding.

Comparison of responses

To contextualize ChatGPT4's performance, its responses were contrasted with those from ChatGPT-3.5 and ChatGPT4-

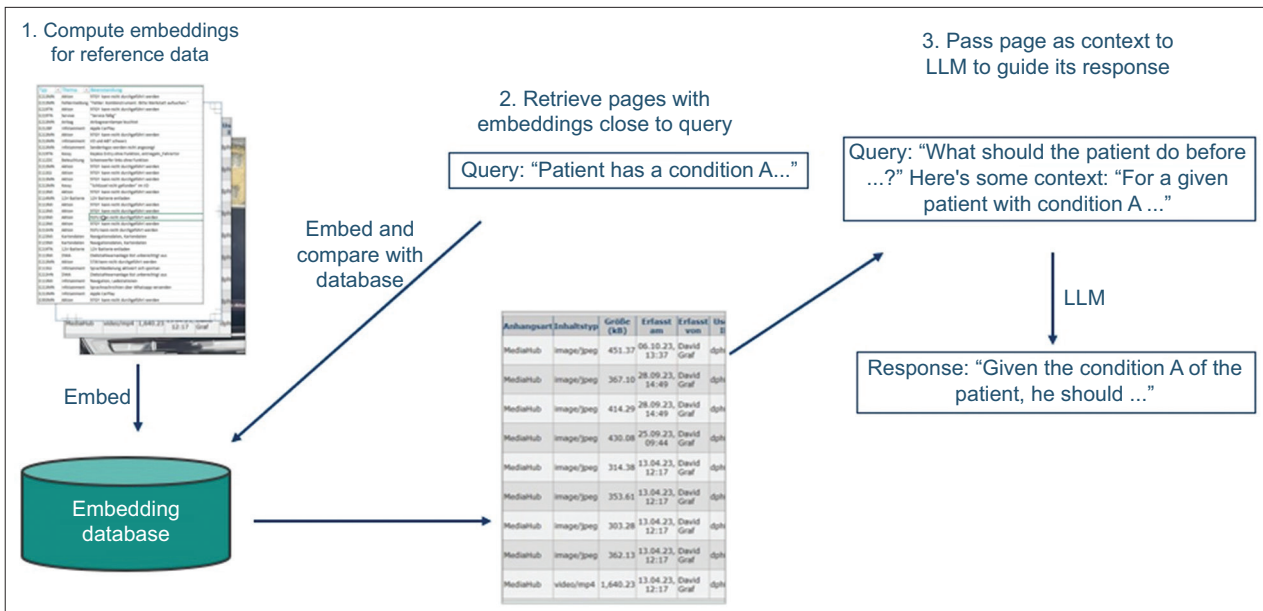


Figure 1 Block diagram explaining RAG
 RAG, retrieval-augmented generation; LLM, large language model

RAG, focusing on accuracy, coherence, and medical relevance. The ChatGPT4-RAG model, known for integrating external information retrieval into the response generation process, presented a unique approach by dynamically pulling data from a corpus to enhance the quality and relevance of its answers [8].

iterations, the essence and content of the responses showed remarkable consistency. Notably, the use of healthcare-specific jargon was minimal.

General perception of gastroenterologists regarding ChatGPT

To gauge the overall perception of gastroenterologists regarding ChatGPT, a 10-item questionnaire was carefully designed and deployed via Google Forms. This digital questionnaire was subsequently disseminated to 10 gastroenterologists through email, ensuring a targeted approach to gather insights from professionals within the field. Upon the completion of data collection, the responses received were thoroughly analyzed to extract meaningful conclusions about the medical community’s stance on the efficacy and reliability of ChatGPT in the context of gastroenterology.

Analysis of ChatGPT responses on anticoagulation for common GI procedures

Nine of 10 gastroenterologists agreed to respond to the questionnaire. The details of their responses are provided in Table 1 and illustrated in Fig. 2.

Results

Diagnostic and therapeutic EGD

For “Diagnostic and Therapeutic EGD,” the responses evaluated show that 51.8% were fully accurate, aligning completely with established medical facts and knowledge. In addition, 22.2% of the answers were considered generally accurate, with minor inaccuracies or omissions, but largely correct. Concerningly, 25.9% of the responses were predominantly or completely inaccurate, deviating significantly from factual medical knowledge.

General evaluation

ERCP with and without stent placement

The complete set of questions and their corresponding response logs have been meticulously documented in the Appendices for comprehensive review. Overall, ChatGPT4 demonstrated a proficient ability to address the queries with clarity and simplicity, utilizing plain English for easy comprehension. Despite variations in phrasing across different

In the “ERCP With and Without Stent Placement” category, 42.8% of ChatGPT’s responses were fully accurate. A slightly higher percentage, 46.0%, was generally accurate, indicating good reliability in the provided information. The proportion of predominantly or completely inaccurate responses was notably low, at 11.1%.

Table 1 Analysis of ChatGPT 4 responses on anticoagulation for common gastrointestinal procedures

Questions	CHATGPT 4 Response	Evaluator 1 (Fellow)	Evaluator 2 (Consultant)	Evaluator 3 (Fellow)	Evaluator 4 (Consultant)	Evaluator 5 (Fellow)	Evaluator 6 (Consultant)	Evaluator 7 (Consultant)	Evaluator 8 (Fellow)	Evaluator 9 (Fellow)
Diagnostic and therapeutic EGD (25.9%:0 – 22.2%:1 – 51.8%:2)										
A patient with atrial fibrillation (CHA ₂ DS ₂ -VASc >=7) is scheduled for diagnostic EGD. When should warfarin be stopped?	Stop warfarin 5 days before EGD	1	0	2	0	2	2	2	2	1
A patient on apixaban with a recent history of stroke or TIA is scheduled for treatment of varices. When should apixaban be stopped?	Discontinue apixaban 48 hours before treatment	2	1	2	1	1	2	2	1	1
What is the protocol for managing a patient on dual antiplatelet therapy with a recent percutaneous intervention and a drug-eluting stent scheduled for diagnostic EGD?	Continue antiplatelet therapy; assess bleeding risk	2	2	0	1	2	0	0	2	2
For a patient with a drug-eluting stent undergoing elective endoscopy, what is the recommended timing for resuming thienopyridine and aspirin therapy?	Resume thienopyridine immediately after endoscopy; continue aspirin	2	0	2	1	2	2	2	2	2
A patient with a bare-metal stent placed 6 months ago on dual antiplatelet therapy is scheduled for a diagnostic endoscopic procedure. How should the antiplatelet therapy be managed?	Continue antiplatelet therapy; prioritize stent protection	2	2	0	1	2	0	0	2	2
In a patient with LVAD on Warfarin, scheduled for diagnostic EGD what is the recommended management of warfarin?	Continue warfarin; aim for the lower end of the therapeutic range	2	1	0	1	0	0	0	0	2
ERCP with or without stent placement (11.1%:0 – 46%:1 – 42.8%:2)										
For a patient with recent VTE (less than 3 months) undergoing ERCP with stent placement, how many days before the procedure should warfarin be discontinued?	Discontinue warfarin 5 days prior to ERCP	1	0	2	1	2	2	2	2	1
For a patient on DOACs with non-severe thrombophilia scheduled for ERCP with stent placement, what adjustments are needed for their anticoagulant therapy?	Stop DOACs 24-48 hours before; resume post-procedure	0	1	1	1	2	1	1	2	0
A patient with a history of atrial fibrillation on dabigatran is scheduled for ERCP. What is the recommended management strategy for dabigatran?	Stop dabigatran 1-2 days before ERCP; resume after assessing bleeding risk	2	1	1	1	2	1	1	2	2

(Contd...)

Table 1 (Continued)

Questions	CHATGPT 4 Response	Evaluator 1 (Fellow)	Evaluator 2 (Consultant)	Evaluator 3 (Fellow)	Evaluator 4 (Consultant)	Evaluator 5 (Fellow)	Evaluator 6 (Consultant)	Evaluator 7 (Consultant)	Evaluator 8 (Fellow)	Evaluator 9 (Fellow)
In a patient with a bi-leaflet aortic valve replacement on warfarin, scheduled for therapeutic ERCP, how should the warfarin be managed?	Stop warfarin 5 days before; bridge with LMWH; resume post-procedure	2	0	2	0	1	2	2	1	2
How should a patient with a mechanical heart valve on warfarin be managed for a scheduled diagnostic endoscopic retrograde cholangiopancreatography (ERCP)?	Stop warfarin 5 days before; bridge if necessary; resume post-ERCP	1	1	2	1	0	2	2	0	1
In a patient on dual antiplatelet therapy after a recent drug-eluting stent placement, scheduled for urgent ERCP for choledocholithiasis, how should the therapy be adjusted?	Continue therapy; consult cardiology for individual risk assessment	2	2	1	1	2	1	1	2	2
In a patient with recent VTE (less than 3 months) on DOAC, scheduled for ERCP with biliary stent placement, what is the recommended management of DOAC	In such cases, temporarily stopping DOAC before ERCP, bridging with low molecular weight heparin, and resuming DOAC post-procedure is recommended	2	1	1	1	2	1	1	2	1
EUS with FNA (14.8%:0 – 50%:1 – 35.1%:2)										
When should therapeutic dose enoxaparin be resumed after EUS with FNA in a patient with a mechanical valve and recent stroke/TIA within 3 months?	Resume enoxaparin 48-72 hours after EUS with FNA	1	1	1	0	0	1	1	0	1
In a patient with a recent drug-eluting stent and on dual antiplatelet therapy, scheduled for EUS with FNA, how should the antiplatelet therapy be managed?	Continue dual antiplatelet therapy; assess risk individually	2	1	0	1	2	0	0	2	2
How should anticoagulation be adjusted for a patient with a history of recurrent VTE on rivaroxaban undergoing EUS with FNA?	Stop rivaroxaban 24-48 hours before EUS-FNA, resume post-procedure	2	1	2	1	2	2	2	2	2

(Contd...)

Table 1 (Continued)

Questions	CHATGPT 4 Response	Evaluator 1 (Fellow)	Evaluator 2 (Consultant)	Evaluator 3 (Fellow)	Evaluator 4 (Consultant)	Evaluator 5 (Fellow)	Evaluator 6 (Consultant)	Evaluator 7 (Consultant)	Evaluator 8 (Fellow)	Evaluator 9 (Fellow)
In a patient with severe thrombophilia on DOAC, scheduled for EUS with FNA what is the recommended management of DOAC	In severe thrombophilia on DOAC, planned for EUS with FNA, consult hematologist; consider temporarily stopping DOAC, bridging with heparin	1	0	1	1	2	1	1	2	1
In a patient with atrial fibrillation with recent history of stroke or TIA, scheduled for EUS with FNA, what is the recommended management of warfarin?	For atrial fibrillation with recent stroke/TIA on warfarin, scheduled for EUS with FNA: temporarily halt warfarin, bridge with heparin	2	0	1	1	2	1	1	2	2
In a patient with non-severe thrombophilia (heterozygous factor V Leiden or prothrombin gene mutation), scheduled for EUS with FNA what is the recommended management of DOAC?	In patients with non-severe thrombophilia undergoing EUS with FNA, temporary discontinuation of DOACs is recommended, with individualized risk assessment	1	1	1	1	2	1	1	2	1
Diagnostic and therapeutic colonoscopy (9%-0-40%;1-50%;2)										
In a patient with a caged ball or tilting disc valve in mitral/aortic position, undergoing polypectomy, what is the guidance for warfarin and enoxaparin management?	Stop warfarin 5 days before, bridge with enoxaparin, resume post-polypectomy	2	0	2	1	2	2	2	2	2
What is the protocol for managing warfarin therapy in a patient with atrial fibrillation (CHA ₂ DS ₂ -VASc >=7) undergoing polypectomy?	Stop warfarin 5 days before; bridge with LMWH if needed	1	0	2	1	2	2	2	2	1
What is the recommended approach for resuming anticoagulation in a patient with a history of DVT undergoing endoscopic mucosal resection?	Resume anticoagulation 48-72 hours post-procedure, based on bleeding risk	2	1	2	1	2	2	2	2	2

(Contd...)

Table 1 (Continued)

Questions	CHATGPT 4 Response	Evaluator 1 (Fellow)	Evaluator 2 (Consultant)	Evaluator 3 (Fellow)	Evaluator 4 (Consultant)	Evaluator 5 (Fellow)	Evaluator 6 (Consultant)	Evaluator 7 (Consultant)	Evaluator 8 (Fellow)	Evaluator 9 (Fellow)
In a patient with a history of atrial fibrillation on apixaban scheduled for sigmoidoscopy, when should apixaban be stopped and restarted?	Stop apixaban 24-48 hours before, resume 24-48 hours after sigmoidoscopy	2	1	1	0	2	1	1	2	2
In a patient with atrial fibrillation (CHA ₂ DS ₂ VASc 1-6), scheduled for diagnostic colonoscopy, what is the recommended management of DOAC?	For atrial fibrillation patients with CHA ₂ DS ₂ VASc 1-6 undergoing colonoscopy, temporarily hold DOACs; timing based on bleeding risk and renal function	1	1	1	0	2	1	1	2	1
In a patient with Recent VTE (less than 3 months), scheduled for diagnostic colonoscopy, what is the recommended management of DOAC?	For patients with recent VTE undergoing colonoscopy, temporarily withhold DOACs, considering individual bleeding and thrombotic risks, and consult a hematologist	1	1	1	0	2	1	1	2	1
PEG Placement (15%:0 – 44%:1 – 17%:2)										
For a patient on dabigatran (CrCl<50) with recent VTE, undergoing PEG placement, what is the appropriate timing for stopping the medication?	Stop dabigatran 2-4 days before PEG placement	2	1	0	1	2	0	0	2	2
For a patient with a mechanical mitral valve without major stroke risk factors undergoing endoscopic submucosal dissection, how should warfarin therapy be altered?	Consult cardiologist; may need bridging anticoagulation	1	1	2	1	2	2	2	2	1
What is the recommendation for managing warfarin therapy in a patient with VTE after 3 months, scheduled for percutaneous endoscopic jejunostomy?	Stop warfarin 5 days before; bridge if necessary	2	0	2	1	1	2	2	1	2
For a patient with a history of DVT on rivaroxaban undergoing PEG, what is the protocol for anticoagulation management?	Stop rivaroxaban 24 hours before PEG; resume post-procedure	1	1	2	1	1	2	2	1	1

(Contd...)

Table 1 (Continued)

Questions	CHATGPT 4 Response	Evaluator 1 (Fellow)	Evaluator 2 (Consultant)	Evaluator 3 (Fellow)	Evaluator 4 (Consultant)	Evaluator 5 (Fellow)	Evaluator 6 (Consultant)	Evaluator 7 (Consultant)	Evaluator 8 (Fellow)	Evaluator 9 (Fellow)
What are the considerations for a patient on rivaroxaban with a history of recurrent VTE undergoing an elective endoscopic procedure for PEG removal?	Stop rivaroxaban 24 hours before; resume post-procedure	2	1	0	1	1	0	0	1	1
In a patient with recent VTE (less than 3 months), scheduled for enteral stent deployment, what is the recommended management of DOAC?	Temporarily switch to low molecular weight heparin (LMWH) before and after stent deployment	1	0	2	1	2	2	2	2	1
In a patient with recent VTE (less than 3 months), scheduled for enteral stent deployment what is the recommended management of warfarin?	Recent VTE, on warfarin: Continue warfarin, adjust dosage based on INR, bridge with LMWH if needed	1	1	0	1	2	0	0	2	1
In a patient with LVAD on warfarin, scheduled for enteral stent deployment what is the recommended management of warfarin?	LVAD, on Warfarin: Maintain therapeutic INR; consult cardiology for specific INR target, bridging strategy	2	1	0	1	2	0	0	2	2
In a patient with mechanical mitral valve with major risk factors (A fib, prior stroke/TIA, prior valve thrombosis, rheumatic heart disease, hypertension, CHF; age >=75), scheduled for enteral stent deployment what is the recommended management of warfarin?	Mitral valve, major risks, on warfarin: Aim for higher INR target; bridge with LMWH; cardiology consultation essential	1	1	1	1	2	1	1	2	1
In a patient with severe thrombophilia on warfarin, scheduled for enteral stent deployment what is the recommended management of warfarin?	Severe thrombophilia, on warfarin: Continue warfarin; monitor INR closely; consult hematology for perioperative management	1	1	0	1	1	0	0	1	1
In a patient with recurrent VTE on DOAC, scheduled for enteral stent deployment what is the recommended management of warfarin?	Recurrent VTE, on DOAC, managing warfarin: Transition to warfarin with bridging anticoagulation, consult thrombosis specialist	1	0	1	1	2	1	1	2	0

Enteral stent deployment (22%:0 – 50%:1 – 27%:2)

EGD, esophagogastroduodenoscopy; TIA, transient ischemic attack; LVAD, left ventricular assist device; ERCP, endoscopic retrograde cholangiopancreatography; VTE, venous thromboembolism; DOAC, direct oral anticoagulant; LMWH, low molecular weight heparin; EUS, endoscopic ultrasound; FNA, fine-needle aspiration; DVT, deep venous thrombosis; PEG, percutaneous endoscopic gastrostomy; INR, international normalized ratio; A fib, atrial fibrillation; CHF, congestive heart failure

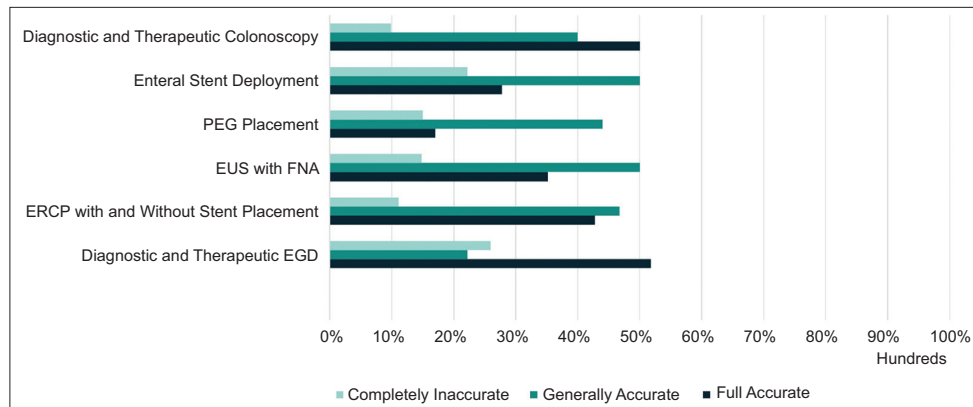


Figure 2 GPT-4 responses on anticoagulation from common gastrointestinal procedures
EGD, esophagogastroduodenoscopy; ERCP, endoscopic retrograde cholangiopancreatography; EUS, endoscopic ultrasound; FNA, fine-needle aspiration; PEG, percutaneous endoscopic gastrostomy

EUS with FNA

Responses related to “EUS with FNA” demonstrated that 35.19% were fully accurate. The majority of the responses, 50.00%, fell into the generally accurate category, suggesting a strong foundation of correct information with some room for improvement. Only 14.81% of answers were identified as predominantly or completely inaccurate.

Diagnostic and therapeutic colonoscopy

For “Diagnostic and Therapeutic Colonoscopy,” the analysis shows that 50.0% of the responses were fully accurate. A significant majority, 41%, were generally accurate, indicating that while the responses are largely on track, minor inaccuracies or omissions exist. Only 9% of the responses were predominantly or completely inaccurate, showcasing a relatively high level of reliability in ChatGPT’s responses to colonoscopy-related questions.

PEG placement

For queries about “PEG Placement,” the fully accurate responses accounted for 17% of the total. The largest share, 44%, was generally accurate, while 15% of the responses were categorized as predominantly or completely inaccurate, indicating a need for careful review of the information provided in this area.

Enteral stent deployment

In the area of “Enteral Stent Deployment,” 27.78% of responses were fully accurate, and a significant 50.00% were generally accurate, reflecting reliable information. However, 22.22% of the responses were predominantly or completely inaccurate, highlighting areas where further accuracy is needed.

Overall feedback on anticoagulation queries addressed by ChatGPT

The percentage of gastroenterologists who agree with the information provided by ChatGPT on anticoagulation, specifically for the question “Do you agree with the ChatGPT-provided information?” was 80%.

Comparison of responses

A comparison of the responses of ChatGPT-3.5, ChatGPT-4, and ChatGPT-4-RAG is provided in Table 2 and Fig. 3.

Accuracy and coherence of ChatGPT-3.5 vs. ChatGPT-4

The comparison reveals a distinct performance gap between ChatGPT-4 and ChatGPT-3.5. ChatGPT-4 showed superior accuracy and coherence, with 7 of 36 responses (19.44%) fully aligning with medical standards and showing high clarity. In contrast, ChatGPT-3.5, while maintaining a level of general accuracy with 4 of 36 responses (11.11%) being fully accurate and 8 of 36 (22.22%) generally accurate, exhibited minor inaccuracies and less clarity in its responses.

Accuracy and coherence of ChatGPT-4 vs. ChatGPT4-RAG

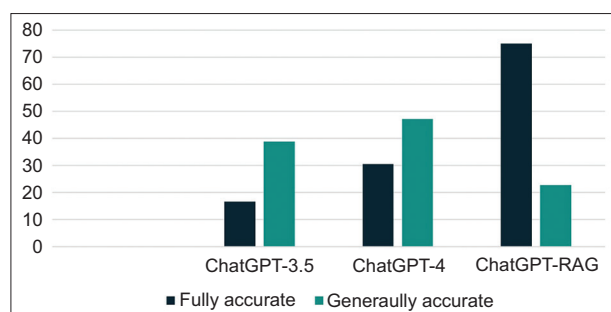
Our analysis of the ChatGPT4-RAG model underscored its exceptional ability to integrate current and specific information, yielding a higher proportion of fully accurate responses—19 of 36 (52.78%)—and demonstrating its potential for more detailed and nuanced responses. Additionally, 4 of 36 responses (11.11%) were generally accurate, indicating RAG’s comprehensive understanding and its edge in delivering clinically relevant advice over traditional models.

Table 2 Comparison of responses of RAG, ChatGPT-4, ChatGPT-3.5

Model	Answers in 2 (Fully accurate)	Answers in 1 (Generally accurate)
RAG	27/36 (75%)	8/36 (22.2%)
GPT-4	11/36 (30.5%)	17/36 (47.2%)
GPT-3.5	6/36 (16.6%)	14/36 (38.9%)

RAG, retrieval-augmented generation

Medical accuracy of each response (Score Key for Tables 1 and 2)
2=Fully accurate: The content aligns completely with established medical facts and knowledge
1=Generally accurate: Although the response might have minor inaccuracies or omissions, it remains largely correct
0=Predominantly or completely inaccurate: The answer deviates significantly or entirely from factual medical knowledge

**Figure 3** Comparison ChatGPT-3.5 vs. ChatGPT-4 vs. ChatGPT-RAG responses

General perception of gastroenterologists regarding ChatGPT

The feedback from gastroenterologists on ChatGPT, collated through a structured questionnaire, revealed insightful perspectives on its utility in gastroenterology. Table 3 summarizes the findings. The analysis highlighted several key findings:

1. Consistency with Latest Guidelines: 60% believe ChatGPT provides information consistent with the latest gastroenterology guidelines.
2. Reduction of Errors in Medication Management: 90% believe that ChatGPT can help reduce errors in medication management, particularly in complex areas like anticoagulation therapy.
3. Reduction of Time Spent on Patient Education: 100% believe that ChatGPT can reduce the time healthcare professionals spend on patient education without compromising quality.
4. Use as a Patient Education Resource: 60% would use ChatGPT as a patient education resource in their practice.
5. Confidence in Maintaining Patient Confidentiality: only 10% are confident in ChatGPT's ability to maintain patient confidentiality and comply with healthcare privacy regulations.
6. Trust in Autonomously Handling Patient Education: 0% would trust ChatGPT to autonomously handle all aspects of patient education.

7. Providing Immediate Responses Outside Clinic Hours: 70% believe ChatGPT can provide immediate responses to patient questions outside of clinic hours, improving access to information.
8. Concerns over Diminishing the Physician–Patient Relationship: 50% are concerned about reliance on AI tools like ChatGPT potentially diminishing the physician–patient relationship.
9. Recommendation for Continuing Medical Education: 70% would recommend ChatGPT as a tool for continuing medical education and professional development in gastroenterology.

Discussion

In an era when the integration of AI into healthcare is rapidly evolving, our study aimed to investigate the effectiveness of ChatGPT4 (with and without RAG) and ChatGPT3.5 in managing anticoagulation queries prior to endoscopic procedures. This study critically evaluated ChatGPT's ability to provide answers to such queries that are accurate, coherent, and medically relevant. Our results show that GPT-based LLMs, through their advanced natural language processing and extensive training datasets, are a promising tool for gastroenterologists, showing a significant improvement in accuracy and coherence compared to their predecessors.

The proficiency of ChatGPT-4 in addressing anticoagulation-related questions with minimal healthcare jargon, and its ability to generate responses that are consistent across various iterations, underscore its potential as a reliable resource for medical professionals. The majority of gastroenterologists expressed confidence in ChatGPT's ability to reduce errors in medication management, particularly in complex areas like anticoagulation therapy, and to serve as an efficient patient education resource, thus potentially enhancing the quality of care and patient safety. The lack of standard guidelines regarding anticoagulation for endoscopic procedures may have led to the inter-observer variance seen in scores. Nevertheless, the accuracy levels of ChatGPT-4's responses concerning Diagnostic and Therapeutic EGD, ERCP With and Without Stent Placement, and other procedures were impressive. For instance, 51.8% of responses related to Diagnostic and Therapeutic EGD were considered fully accurate, and 50% of those for Diagnostic and Therapeutic Colonoscopy were generally accurate, demonstrating the nuanced understanding of ChatGPT in these medical contexts. Overall agreement with the information provided by ChatGPT among gastroenterologists was 80%, indicating strong confidence in the utility of ChatGPT. Furthermore, 90% of respondents believe ChatGPT can help reduce medication errors, while 100% believe it can reduce the amount of time spent on patient education. These statistics underscore the potential of ChatGPT to improve medical practice and patient care significantly [11,12].

The evolution from GPT-3, introduced in June 2020, through ChatGPT-3.5, to the unveiling of ChatGPT-4 in November 2022, marks a trajectory of significant advances in natural language processing and understanding [13,14]. Each iteration has shown

Table 3 General perception of gastroenterologists regarding ChatGPT

Does ChatGPT provide information consistent with the latest gastroenterology guidelines	Do you believe that ChatGPT can help reduce errors in medication management, particularly in complex cases like anticoagulation therapy?	Can ChatGPT reduce the time healthcare professionals spend on patient education without compromising quality?	Would you use ChatGPT as a patient education resource in your practice?	Are you confident in ChatGPT's ability to maintain patient confidentiality and comply with healthcare privacy regulations?	Would you trust ChatGPT to autonomously handle all aspects of patient education?	Can ChatGPT provide immediate responses to patient questions outside of clinic hours, improving access to information?	Are you concerned about the reliance on AI tools like ChatGPT potentially diminishing the physician-patient relationship?	Would you recommend ChatGPT as a tool for continuing medical education and professional development in gastroenterology?
No	Yes	Yes	No	No	No	Yes	No	No
No	Yes	Yes	No	No	No	Yes	Yes	Yes
Yes	Yes	Yes	Yes	No	No	Yes	Yes	Yes
Yes	Yes	Yes	Yes	No	No	No	Yes	Yes
Yes	No	Yes	Yes	No	No	Yes	No	Yes
Yes	Yes	Yes	No	No	No	Yes	Yes	No
Yes	Yes	Yes	Yes	No	No	Yes	No	Yes
No	Yes	Yes	Yes	No	No	No	No	No
No	Yes	Yes	Yes	No	No	No	Yes	Yes
Yes	Yes	Yes	No	Yes	No	Yes	No	Yes

substantial improvements in generating contextually relevant and coherent responses. Notably, ChatGPT-4 demonstrates superior performance, particularly in delivering medically accurate, concise, and contextually relevant responses, with a full accuracy rate of 30.5% and general accuracy of 47.2%, compared to ChatGPT-3.5's 16.6% and 38.9%, respectively. This leap in accuracy can be attributed to ChatGPT-4's advanced training, refined algorithms, and the integration of the RAG model, which further enhances its capability by achieving 75% in full accuracy, showcasing its exceptional ability to incorporate up-to-date, specific information, which is especially useful in complex medical scenarios [15,16]. Unlike its predecessors, which were trained on internet datasets up to their release, ChatGPT-4 relies on its extensive dataset up to its last update and leverages reinforcement learning from human feedback to refine its output, ensuring high-quality responses without real-time internet access.

However, as in findings from the previous literature [17-19], concerns regarding patient confidentiality and the preservation of the physician-patient relationship in the context of increasing AI utilization highlight the need for careful integration of such technologies into clinical practice. Only 10% of respondents are confident in ChatGPT's ability to maintain patient confidentiality and comply with healthcare privacy regulations, indicating significant data security concerns. In addition, none of the gastroenterologists surveyed would trust ChatGPT to autonomously handle all aspects of patient education, reflecting skepticism about AI's ability to manage nuanced patient interactions without human oversight. Additionally, 50% of respondents expressed concern that reliance on AI tools such as ChatGPT potentially diminishes the physician-patient relationship. This suggests a perception that, while AI can

improve certain aspects of care, it cannot replace the essential human elements, such as empathy, understanding and trust, that define the medical profession. These findings underscore the importance of balancing technological advancement with the intrinsic values that define patient care and ensuring that AI serves as a complement to, rather than a replacement for, the critical human elements of medical practice.

One of the key strengths of our study is its focus on the latest version of ChatGPT, which provides insights into how AI is at the forefront of solving medical issues, alongside the remarkable addition of the RAG model, demonstrating its ability to elevate the accuracy and depth of responses through data retrieval integration. The inclusion of a diverse group of gastroenterologists at different stages of their careers offers a broad perspective on the clinical impact of AI. However, this study had its limitations. Our analysis did not extend to other AI resources, such as BLOOM, LaMBDA/Bard or LLaMA, which represent significant advances in AI research and application. Furthermore, given that 2 of the 3 potential options for evaluators were favorable towards ChatGPT, the survey results might be biased towards positive findings. Additionally, the subjectivity of the assessments is evident, with 1 evaluator rating only 3 of 36 responses as "Fully accurate," while another rated 27 of 36 as "Fully accurate." This variability, along with differences in evaluator experience (fellows vs. consultants), is a limitation. Moreover, the questionnaire used in this study lacks validation highlighting the need for more standardized assessment tools. Although previous studies have shown that ChatGPT-4 outperformed ChatGPT-3.5 and Bard [20,21], the rapid development of AI technologies also suggests that our results may need to be re-evaluated as new versions of

ChatGPT and other AI models are developed that may provide even more sophisticated tools for healthcare professionals. Nevertheless, the inclusion of RAG in our study underscores the potential of retrieval-augmented models in enhancing the utility and applicability of AI in healthcare, pointing towards a future where AI can offer more personalized, accurate and comprehensive medical advice [22].

In summary, our study supports the integration of AI tools such as ChatGPT in gastroenterology and advocates their role in improving care. However, it also underscores the critical need for ongoing evaluation to ensure these technologies complement the essential human elements of medical practice. While ChatGPT4-RAG demonstrates significant advances over previous AI versions, medicolegal considerations dictate that a 75% full accuracy rate is insufficient for independent clinical decision-making. AI should serve to support and confirm clinician decisions rather than generate them. Future research should address the identified limitations, particularly in enhancing data privacy and understanding the long-term impact of AI on the physician–patient relationship.

Summary Box

What is already known:

- Managing anticoagulation in patients undergoing gastrointestinal (GI) procedures is complex, raising questions about safety, timing, and management
- Artificial intelligence (AI), particularly large language models like ChatGPT, shows the potential to provide reliable medical guidance for anticoagulation management
- Challenges in AI applications include inaccuracies, such as generating misleading information, that are due to the model's inherent training limitations

What the new findings are:

- ChatGPT-4, enhanced by retrieval-augmented generation, demonstrates 75% full accuracy in answering anticoagulation-related questions, markedly higher than previous versions (ChatGPT-3.5)
- Survey results from gastroenterologists indicate that 90% believe ChatGPT can reduce errors in complex medication management scenarios, while 100% acknowledge its potential to decrease the time spent on patient education
- Despite technological advances, only 10% of surveyed gastroenterologists are confident in ChatGPT's ability to maintain patient confidentiality, highlighting ongoing concerns about data security in AI applications within healthcare

References

1. Acosta RD, Abraham NS, Chandrasekhara V, et al; ASGE Standards of Practice Committee. The management of antithrombotic agents for patients undergoing GI endoscopy. *Gastrointest Endosc* 2016;**83**:3-16.
2. Maida M, Sferrazza S, Maida C, et al. Management of antiplatelet or anticoagulant therapy in endoscopy: a review of literature. *World J Gastrointest Endosc* 2020;**12**:172-192.
3. Scridon A, Balan AI. Challenges of anticoagulant therapy in atrial fibrillation-focus on gastrointestinal bleeding. *Int J Mol Sci* 2023;**24**:6879.
4. Veitch AM, Vanbiervliet G, Gershlick AH, et al. Endoscopy in patients on antiplatelet or anticoagulant therapy, including direct oral anticoagulants: British Society of Gastroenterology (BSG) and European Society of Gastrointestinal Endoscopy (ESGE) guidelines. *Gut* 2016;**65**:374-389.
5. Klang E, Sourash A, Nadkarni GN, Sharif K, Lahat A. Evaluating the role of ChatGPT in gastroenterology: a comprehensive systematic review of applications, benefits, and limitations. *Therap Adv Gastroenterol* 2023;**16**:17562848231218618.
6. Bekbolatova M, Mayer J, Ong CW, Toma M. Transformative potential of AI in healthcare: definitions, applications, and navigating the ethical landscape and public perspectives. *Healthcare (Basel)* 2024;**12**:125.
7. Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. *N Engl J Med* 2023;**388**:1233-1239.
8. Lewis P, Perez E, Piktus A, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. 2021. Available from: <http://arxiv.org/abs/2005.11401> [Accessed 12 July 2024].
9. Workshop B, Scao TL, et al. BLOOM: a 176B-parameter open-access multilingual language model. 2022. Available from: <https://arxiv.org/abs/2211.05100> [Accessed 12 July 2024].
10. Gibney E. Open-source language AI challenges big tech's models. *Nature* 2022;**606**:850-851.
11. Wójcik S, Rulkiewicz A, Pruszczyk P, Lisik W, Poboży M, Domienik-Karłowicz J. Beyond ChatGPT: what does GPT-4 add to healthcare? The dawn of a new era. *Cardiol J* 2023;**30**:1018-1025.
12. Moore S. What does ChatGPT mean for healthcare? Available from: <https://www.news-medical.net/health/What-does-ChatGPT-mean-for-Healthcare.aspx> [Accessed 12 July 2024].
13. Assaraf N. OpenAI's ChatGPT: Optimizing language models for dialogue – cloudHQ. 2022. Available from: <https://blog.cloudhq.net/openai-chatgpt-optimizing-language-models-for-dialogue/> [Accessed 12 July 2024].
14. Brown TB, Mann B, Ryder N, et al. Language models are few-shot learners. 2020. Available from: <http://arxiv.org/abs/2005.14165> [Accessed 12 July 2024].
15. Massey PA, Montgomery C, Zhang AS. Comparison of ChatGPT-3.5, ChatGPT-4, and orthopaedic resident performance on orthopaedic assessment examinations. *J Am Acad Orthop Surg* 2023;**31**:1173-1179.
16. Knoedler L, Alfertshofer M, Knoedler S, et al. Pure wisdom or potemkin villages? A comparison of ChatGPT 3.5 and ChatGPT 4 on USMLE step 3 style questions: quantitative analysis. *JMIR Med Educ* 2024;**10**:e51148.
17. Wang C, Liu S, Yang H, Guo J, Wu Y, Liu J. Ethical considerations of using ChatGPT in health care. *J Med Internet Res* 2023;**25**:e48009.
18. The Lancet Digital Health. ChatGPT: friend or foe? *Lancet Digit Health* 2023;**5**:e102.
19. Meskó B, Topol EJ. The imperative for regulatory oversight of large language models (or generative AI) in healthcare. *NPJ Digit Med* 2023;**6**:120.
20. Tariq R, Malik S, Khanna S. Evolving landscape of large language models: an evaluation of ChatGPT and Bard in answering patient

- queries on colonoscopy. *Gastroenterology* 2024;**166**:220-221.
21. Patil NS, Huang RS, van der Pol CB, Larocque N. Comparative performance of ChatGPT and Bard in a text-based radiology knowledge assessment. *Can Assoc Radiol J* 2024;**75**:344-350.
 22. Bohr A, Memarzadeh K. Chapter 2 - The rise of artificial intelligence in healthcare applications. In: Bohr A, Memarzadeh K (Eds.). *Artificial Intelligence in Healthcare*. Academic Press, 2020, pp. 25-60.