Evaluating ChatGPT's Performance in Responding to Questions About Endoscopic Procedures for Patients

Hassam Ali, MD1, Pratik Patel, MD2, Itegbemie Obaitan, MD3,Babu P. Mohan, MD, MS4, Amir Humza Sohail, MD, MSc5, Lucia Smith-Martinez, MD6, Karrisa Lambert, MD1, Manesh Kumar Gangwani, MD7, Jeffrey J. Easler, MD3 , Douglas G. Adler MD, FASGE8

1. Department of Gastroenterology, ECU health medical center/Brody School of Medicine, Greenville, North Carolina, USA
2. Department of Gastroenterology, Mather Hospital/Hofstra University Zucker School of Medicine, Port Jefferson, New York, USA
3. Division of Gastroenterology and Hepatology, Indiana University, Indianapolis, Indiana, USA
4. Department of Gastroenterology & Hepatology, University of Utah School of Medicine, Salt Lake City, Utah, USA.
5. Department of Surgery, NYU Langone Health, Long Island, New York, United States
6. Department of Psychiatry, ECU health medical center/Brody School of Medicine, Greenville, North Carolin, USA
7. Department of Medicine, University of Toledo Medical Center, Toledo, OH,USA
8. Center for Advanced Therapeutic Endoscopy, Porter Adventist Hospital, Centura Health, Denver, Colorado, USA.

Author contributions:

Hassam Ali, Pratik Patel, Itegbemie Obaitan, Babu P. Mohan, Amir Humza Sohail: Conceptualization, Methodology, Software, Data curation, Validation, Writing- Original draft preparation.

Lucia Smith-Martinez, MD6, Karrisa Lambert, MD1, Manesh Kumar Gangwani, MD7, Jeffrey J. Easler: Writing- Reviewing and Editing, Project administration

Babu P. Mohan, Douglas G. Adler: Writing- Reviewing and Editing, Supervision.

Conflicts of interest/ Disclosures:

The authors certify that they have NO affiliations with or involvement in any organization or entity with any financial interest. The authors declare that they have no conflicts of interest.

Ethical statement:

Institutional IRB approval was not obtained for this study as this study does not use any human subjects.

Patient Consent Statement: Not applicable

Abstract

Background and aims: We aimed to assess the accuracy, completeness, and consistency of ChatGPT's responses to frequently asked questions concerning the management and care of patients receiving endoscopic procedures and to compare its performance to Generative Pre-trained Transformer 4 (GPT-4) in providing emotional support.

Methods: Frequently asked questions (N = 117) about esophagogastroduodenoscopy (EGD), colonoscopy, endoscopic ultrasound (EUS), and endoscopic retrograde cholangiopancreatography (ERCP) were collected from professional societies, institutions, and social media. ChatGPT's responses were generated and graded by board-certified gastroenterologists and advanced endoscopists. Emotional support questions were assessed by a psychiatrist.

Results: ChatGPT demonstrated high accuracy in answering questions about EGD (94.8% comprehensive or correct but insufficient), colonoscopy (100% comprehensive or correct but insufficient), ERCP (91% comprehensive or correct but insufficient), and EUS (87% comprehensive or correct but insufficient). No answers were deemed entirely incorrect (0%). Reproducibility was significant across all categories. ChatGPT's emotional support performance was inferior to the newer GPT-4 model.

Conclusion: ChatGPT provides accurate and consistent responses to patient questions about common endoscopic procedures and demonstrates potential as a supplementary information resource for patients and healthcare providers.

Keywords: Artificial Intelligence, Gastrointestinal Endoscopy, Patient Education, Health Communication, Large language models.

Introduction

Endoscopy remains vital in managing gastrointestinal diseases; it provides essential diagnostic and therapeutic interventions for innumerable gastrointestinal conditions. [1] In 2020, approximately 20 million endoscopic procedures were performed in the United States, highlighting the extensive reliance on these procedures in clinical settings. [2]

The most common endoscopic procedures include esophagogastroduodenoscopy (EGD), colonoscopy, endoscopic ultrasound (EUS), and endoscopic retrograde cholangiopancreatography (ERCP). Due to the widespread nature of gastrointestinal disorders and the significance of endoscopic procedures in addressing these issues, patients will frequently have questions regarding these techniques.

Artificial intelligence (AI) has made remarkable strides in natural language processing (NLP) in recent years. [3] Models such as ChatGPT (Generative Pre-trained Transformer) and GPT-4 (Generative Pre-trained Transformer-4), developed by OpenAI, an artificial intelligence research organization based in San Francisco, California, have demonstrated potential for various healthcare applications. [4] These models have been utilized in tasks including responding to medical student examination queries, creating basic medical reports, and offering information on various health-related subjects. [5, 6, 7] ChatGPT could potentially serve as a supplementary information resource for patients, improving patient education and outcomes. [8] Nevertheless, concerns persist about ChatGPT's ability to provide accurate and comprehensive responses to detailed medical questions.

No current literature specifically investigates ChatGPT's capabilities in addressing questions concerning common endoscopic procedures. Our study aims to assess the precision, comprehensiveness, and reliability of ChatGPT's answers to common queries about patient care and management regarding endoscopic procedures. Moreover, we will compare the performance of ChatGPT (freely accessible) to GPT-4 (paid subscription/limited access) when responding to questions posed by patients, as this comparison could reveal further insight into its potential role as a virtual assistant for patients and healthcare providers in the realm of endoscopic procedures.

Methods

Data source

FAQs on endoscopic procedures were sourced from professional societies and institutional websites (Supplemental file 1). Excluding repetitive or unclear questions, we curated 117 on endoscopic procedures (EGD, colonoscopy, EUS, ERCP; Supplemental Table 1-4). Additionally, we tested ChatGPT and GPT-4's psychological support capability with 16 questions (Supplemental Table 5), evaluated by a certified psychiatrist (LSM).

Response generation

ChatGPT, an enhanced GPT-3.5 model introduced in November 2022, incorporates user feedback and restrictions for safer and relevant responses. Two authors input questions twice separately into ChatGPT's March 23 version for reproducibility [8]. GPT-4, a subscription-based model, was not utilized for primary analysis [4].

Grading of questions

Responses for EGD/Colonoscopy were graded by certified gastroenterologists (BPM and PP), while those for EUS/ERCP by advanced endoscopists (IO and JE). A grading system assessed response accuracy and comprehensiveness, with a third reviewer (KL) resolving discrepancies [8].

Emotional support questions and responses

ChatGPT's performance on emotional support was examined using modified FAQs (Supplemental Table 5). The responses were graded by a certified psychiatrist (LSM), ranging from "Not Comprehensive" to "Extremely Comprehensive".

Statistical analysis

Reproducibility, measured by the uniformity of two similarly-graded responses, was evaluated. Disparate responses were graded, with those falling into separate grading groups deemed significantly different. ChatGPT's performance on emotional support questions was compared to GPT-4. Grading proportions for each endoscopic procedure domain were calculated as percentages (N%). Analysis used STATA (version 16.1).

The methodology has been described in detail in Supplementary file 1.

Results

ChatGPT displayed high levels of precision when answering questions about EGD (N =39), colonoscopy (N =22), ERCP (N = 28), and EUS (N = 28) about treatment, lifestyle/aftercare, basic knowledge, and others (Figure 1).

Frequently asked questions about EGD.

The percentage of answers considered comprehensive or correct but insufficient was 94.8% or above for the "basic knowledge," "treatment," "lifestyle/aftercare," and "others" categories (Supplementary Table 1). No answers from ChatGPT were deemed entirely incorrect.

Reproducibility was significant across all categories, with 100% (39/39) of all questions generating two comparable answers. The Reproducibility within specific categories is displayed in Table 1.

Frequently asked questions about Colonoscopy.

The percentage of answers considered comprehensive or correct but insufficient was 100 % or above for the "basic knowledge," "treatment," "lifestyle/aftercare," and "others" categories (Supplementary Table 2). No answers from ChatGPT were deemed entirely incorrect. Reproducibility was significant across all categories, with 95.4% (21/22) of all questions generating two comparable answers. The Reproducibility within specific categories is displayed in Table 1.

Frequently asked questions about ERCP.

The percentage of answers considered comprehensive or correct but insufficient was 91% or above for the "basic knowledge," "treatment," "lifestyle/aftercare," and "others" categories (Supplementary Table 3). No answers from ChatGPT were deemed entirely incorrect. Reproducibility was significant across all categories, with 89.2% (25/28) of all questions generating two comparable answers. The Reproducibility within specific categories is displayed in Table 1.

Frequently asked questions about EUS.

The percentage of answers considered comprehensive or correct but insufficient was 87% or above for the "basic knowledge," "treatment," "lifestyle/aftercare," and "others" categories (Supplementary Table 4). No answers from ChatGPT were deemed entirely incorrect. Reproducibility was significant across all categories, with 92.8% (26/28) of all questions generating two comparable answers. The reproducibility within specific categories is displayed in Table 1.

Emotional support questions about endoscopic procedures

The responses to emotional support questions were graded from 1-5 based on the level of comprehensiveness (Figure 1). Both LLMs (ChatGPT and GPT-4) performed adequately, with all responses being moderate to extremely comprehensive. GPT-4 outperformed ChatGPT responses to emotional questions (Supplementary Table 5). No answers from either LLM were deemed noncomprehensive.

Discussion

This study assessed the precision and consistency of ChatGPT in addressing patient inquiries about endoscopic procedures. ChatGPT generated accurate and relevant responses to these procedures and provided comprehensive information to patients, performing better with basic procedures like EGD and colonoscopy compared to EUS/ERCP. ChatGPT also effectively addressed emotional concerns, showcasing empathy and understanding [15].

Global differences in endoscopic guidelines weren't evaluated. GPT-4 was not examined due to accessibility constraints. ChatGPT could be a supplementary patient resource, enhancing gastroenterological procedure comprehension.

Health literacy is vital for patients undergoing GI endoscopic procedures [9,10]. Despite the need for accessible and accurate information, obtaining easy-to-understand resources can be a struggle. ChatGPT can address this issue by delivering health information conversationally, simplifying complex medical jargon [11, 12], potentially leading to better patient understanding [13]. ChatGPT can support healthcare providers by generating responses to routine patient inquiries, potentially saving time for more complex cases. The accuracy of the responses varies, and with technological improvements, this could increase, possibly boosting provider productivity [14]. ChatGPT and GPT-4 showed empathetic responses to emotional questions. Further research is needed to evaluate this capability, including comprehension of complex inquiries and cultural adaptation.

This study's main strengths include the comprehensive collection of inquiries from authoritative sources. However, ChatGPT has limitations. A few questions got comprehensive responses, hinting at its role as a supplemental tool rather than a replacement for healthcare providers. Discrepancies (<25%) among reviewers demonstrate variation in expert opinions. ChatGPT's training data, limited to 2021, may lead to outdated responses. Its training data's quality remains under review, affecting reliability. Furthermore, ChatGPT struggled with specifics like lab cut-offs or treatment durations. Reviewers' awareness of ChatGPT might have led to stricter grading, potentially underestimating its performance. Finally, globally varying guidelines could lead to confusion or harm if not correctly specified. Further refinement is needed to enhance data reliability and specificity.

ChatGPT can augment healthcare providers, assisting patients with pertinent questions. Our study examined the accuracy and reproducibility of ChatGPT's responses to common patient inquiries on GI endoscopic procedures. ChatGPT frequently provided accurate, albeit sometimes incomplete, responses. The model's advice, varying across regions, should not be solely trusted.

References:

1. ASGE Standards of Practice Committee, Early DS, Ben-Menachem T, et al. Appropriate use of GI endoscopy. *Gastrointest Endosc*. 2012;75(6):1127-1131.
2. Delegge MH. The difficult-to-sedate patient in the endoscopy suite. *Gastrointest Endosc Clin N Am*. 2008;18(4):679-693, viii.
3. Nadkarni PM, Ohno-Machado L, Chapman WW. Natural language processing: an introduction. *J Am Med Inform Assoc*. 2011;18(5):544-551.
4. OpenAI OpenAI: Models GPT-3.5. [(accessed on 23 March 2023)]. Available online: https://chat.openai.com/
5. Ouyang L, Wu J, Jiang X, et al. Training language models to follow instructions with human feedback. *arXiv:220302155 [cs]*. Published online March 4, 2022.
6. Gilson A, Safranek CW, Huang T, et al. How does chatgpt perform on the united states medical licensing examination? The implications of large language models for medical education and knowledge assessment. *JMIR Med Educ*. 2023;9:e45312.
7. Jeblick K, Schachtner B, Dexl J, et al. Chatgpt makes medicine easy to swallow: an exploratory case study on simplified radiology reports. *arXiv:221214882 [cs]*. Published online December 30, 2022.
8. Yeo YH, Samaan JS, Ng WH, et al. Assessing the performance of ChatGPT in answering questions regarding cirrhosis and hepatocellular carcinoma. *Clin Mol Hepatol*. Published online March 22, 2023.
9. Smith SG, von Wagner C, McGregor LM, et al. The influence of health literacy on comprehension of a colonoscopy preparation information leaflet. *Dis Colon Rectum*. 2012;55(10):1074-1080.
10. Kolb JM, Chen M, Tavakkoli A, et al. Patient knowledge, risk perception, and barriers to barrett's esophagus screening. *Am J Gastroenterol*. 2023;118(4):615-626.
11. Morahan-Martin JM. How internet users find, evaluate, and use online health information: a cross-cultural review. *Cyberpsychol Behav*. 2004;7(5):497-510.
12. Zeng QT, Kogan S, Plovnick RM, Crowell J, Lacroix EM, Greenes RA. Positive attitudes and failed queries: an exploration of the conundrums of consumer health information retrieval. *Int J Med Inform*. 2004;73(1):45-55.
13. Miner AS, Laranjo L, Kocaballi AB. Chatbots in the fight against the COVID-19 pandemic. *NPJ Digit Med*. 2020;3:65.
14. Xu L, Sanders L, Li K, Chow JCL. Chatbot for health care and oncology applications using artificial intelligence and machine learning: systematic review. *JMIR Cancer*. 2021;7(4):e27850.
15. Ayers JW, Poliak A, Dredze M, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA Intern Med*. Published online April 28, 2023:e231838.

Table 1. Percentage of questions with significantly different between the two responses.

	Basic Knowledge	Treatment	Lifestyle/aftercare	Other
ERCP	1/12 (8.4%)	1/7 (14.3%)	1/6 (16.7%)	0/3 (0%)
EUS	1/8 (12.5%)	0/9 (0%)	0/4 (0%)	1/7 (14.3%)
EGD	0/11 (0%)	0/9 (0%)	0/9 (0%)	0/10 (0%)
Colonoscopy	1/8 (12.5%)	0/5 (0%)	0/5 (0%)	0/4 (0%)

*Differences between the two responses were assessed by the authors as a binary yes/no answer

Figure 1: Grade of responses by the ChatGPT language model to questions related to endoscopic procedures.

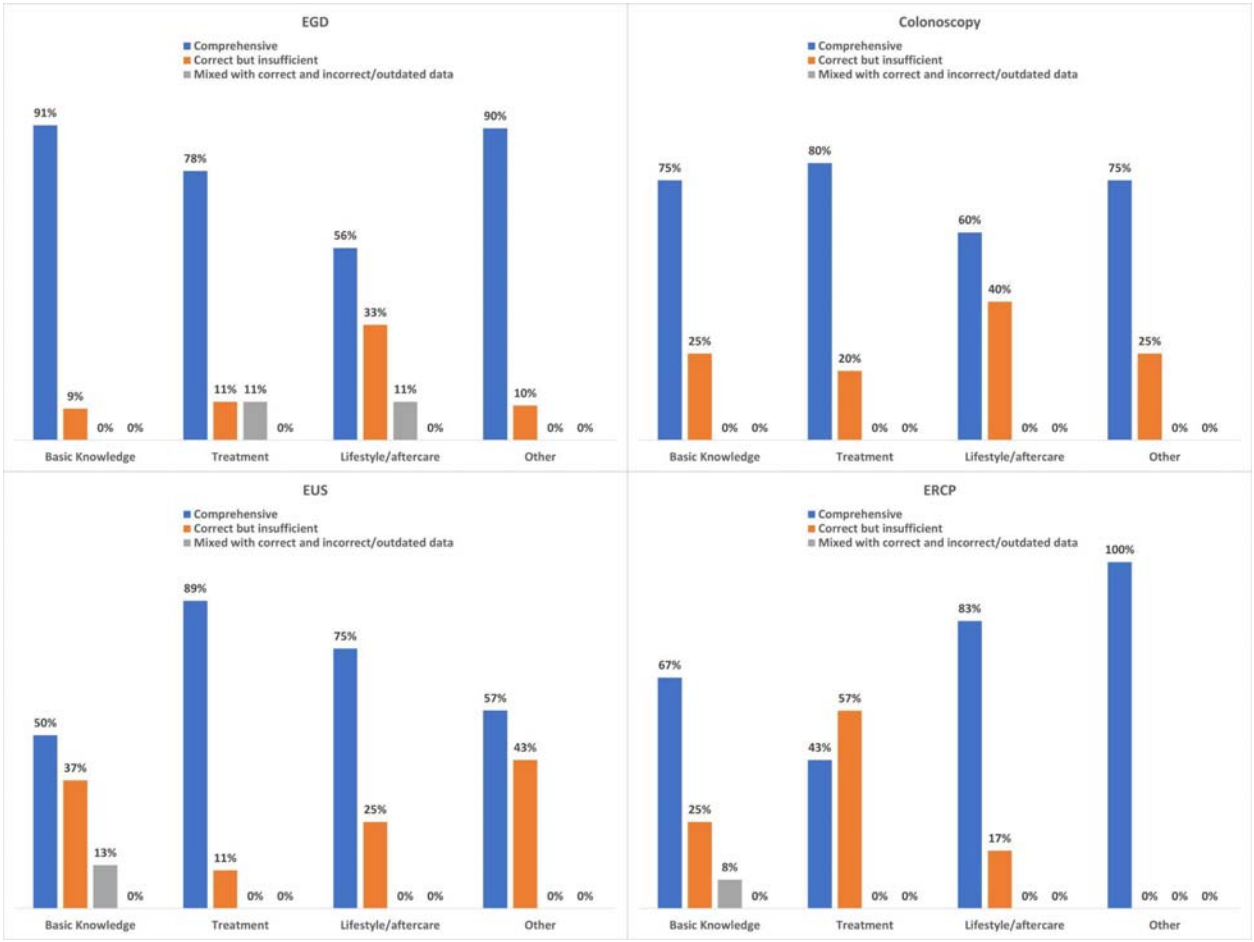
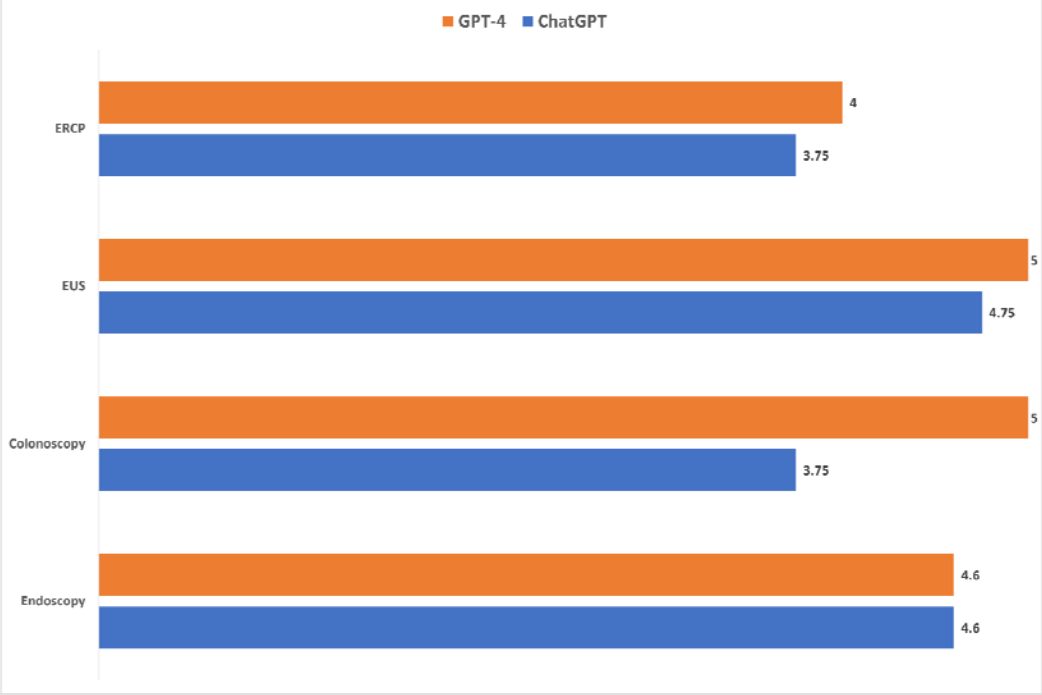


Figure 2: Responses to emotional support statements by ChatGPT and GPT-4 language models



Not Comprehensive: The answer provided minimal information or needed to address the question adequately.
Somewhat Comprehensive: The answer provided some information but left out important details.
Moderately Comprehensive: The answer covered most aspects of the question but may have lacked depth or specificity in certain areas.
Very Comprehensive: The answer addressed all aspects of the question and provided detailed information.
Extremely Comprehensive: The answer was exhaustive, covering all aspects of the question with great detail and specificity.