ISSN 2644-3236 (online)

# Artificial Intelligence in *Gastroenterology*

Quarterly Volume 5 Number 1 April 30, 2024

Published by Baishideng Publishing Group Inc

A I G

Artificial Intelligence in Gastroenterology

Contents

Quarterly Volume 5 Number 1 April 30, 2024

## EDITORIAL



## ORIGINAL ARTICLE

### Observational Study

*Artificial Intelligence in Gastroenterology*

**Quarterly Volume 5 Number 1 April 30, 2024**

# Contents

## ABOUT COVER

Editorial Board Member of *Artificial Intelligence in Gastroenterology*, Rogerio Serafim Parra, MD, PhD, Assistant Professor, Department of Surgery and Anatomy, Ribeirao Preto Medical School, University of Sao Paulo, Ribeirão Preto 14048-900, São Paulo, Brazil. rsparra@hcrp.usp.br

## AIMS AND SCOPE

The primary aim of *Artificial Intelligence in Gastroenterology* (*AIG*, *Artif Intell Gastroenterol*) is to provide scholars and readers from various fields of artificial intelligence in gastroenterology with a platform to publish high-quality basic and clinical research articles and communicate their research findings online.

*AIG* mainly publishes articles reporting research results obtained in the field of artificial intelligence in gastroenterology and covering a wide range of topics, including artificial intelligence in gastrointestinal cancer, liver cancer, pancreatic cancer, hepatitis B, hepatitis C, nonalcoholic fatty liver disease, inflammatory bowel disease, irritable bowel syndrome, and *Helicobacter pylori* infection.

## INDEXING/ABSTRACTING

The *AIG* is now abstracted and indexed in Reference Citation Analysis, China Science and Technology Journal Database.

## RESPONSIBLE EDITORS FOR THIS ISSUE



**NAME OF JOURNAL**
*Artificial Intelligence in Gastroenterology*

**ISSN**
ISSN 2644-3236 (online)

**LAUNCH DATE**
July 28, 2020

**FREQUENCY**
Quarterly

**EDITORS-IN-CHIEF**
Rajvinder Singh, Ferruccio Bonino

**EDITORIAL BOARD MEMBERS**
https://www.wjgnet.com/2644-3236/editorialboard.htm

**PUBLICATION DATE**
April 30, 2024

**COPYRIGHT**
© 2024 Baishideng Publishing Group Inc

**INSTRUCTIONS TO AUTHORS**
https://www.wjgnet.com/bpg/gerinfo/204

**GUIDELINES FOR ETHICS DOCUMENTS**
https://www.wjgnet.com/bpg/GerInfo/287

**GUIDELINES FOR NON-NATIVE SPEAKERS OF ENGLISH**
https://www.wjgnet.com/bpg/gerinfo/240

**PUBLICATION ETHICS**
https://www.wjgnet.com/bpg/GerInfo/288

**PUBLICATION MISCONDUCT**
https://www.wjgnet.com/bpg/gerinfo/208

**ARTICLE PROCESSING CHARGE**
https://www.wjgnet.com/bpg/gerinfo/242

**STEPS FOR SUBMITTING MANUSCRIPTS**
https://www.wjgnet.com/bpg/GerInfo/239

**ONLINE SUBMISSION**
https://www.f6publishing.com

Artificial Intelligence in Gastroenterology

Submit a Manuscript: https://www.f6publishing.com

*Artif Intell Gastroenterol* 2024 April 30; 5(1): 90503

DOI: 10.35712/aig.v5.i1.90503

ISSN 2644-3236 (online)

ORIGINAL ARTICLE

Observational Study

# Evaluating the accuracy and reproducibility of ChatGPT-4 in answering patient questions related to small intestinal bacterial overgrowth

Lauren Schlussel, Jamil S Samaan, Yin Chan, Bianca Chang, Yee Hui Yeo, Wee Han Ng, Ali Rezaie

**Specialty type:** Gastroenterology and hepatology

**Provenance and peer review:** Unsolicited article; Externally peer reviewed.

**Peer-review model:** Single blind

**Peer-review report's classification**

**Scientific Quality:** Grade C, Grade C, Grade C, Grade D, Grade D

**Novelty:** Grade A, Grade B, Grade B, Grade B, Grade B

**Creativity or Innovation:** Grade A, Grade B, Grade B, Grade B, Grade B

**Scientific Significance:** Grade B, Grade B, Grade B, Grade B, Grade C

**P-Reviewer:** Caboclo JLF, Brazil; Wu L, China; Yu YB, China; Zhang C, China

**Received:** December 7, 2023

**Revised:** March 27, 2024

**Accepted:** April 16, 2024

**Published online:** April 30, 2024

**Lauren Schlussel, Jamil S Samaan, Yin Chan, Bianca Chang, Yee Hui Yeo, Ali Rezaie,** Division of Gastroenterology and Hepatology, Cedars-Sinai Medical Center, Los Angeles, CA 90048, United States

**Wee Han Ng,** Bristol Medical School, University of Bristol, BS8 1TH, Bristol, United Kingdom

**Ali Rezaie,** Medically Associated Science and Technology Program, Cedars-Sinai Medical Center, Los Angeles, CA 90048, United States

**Corresponding author:** Ali Rezaie, MD, MSc, FRCPC, Medical Director, Medically Associated Science and Technology Program, Cedars-Sinai Medical Center, Cedars-Sinai, 8730 Alden Drive, Thalians Bldg, #E240, Los Angeles, CA 90048, United States. Ali.rezaie@cshs.org

## Abstract

### *BACKGROUND*

Small intestinal bacterial overgrowth (SIBO) poses diagnostic and treatment challenges due to its complex management and evolving guidelines. Patients often seek online information related to their health, prompting interest in large language models, like GPT-4, as potential sources of patient education.

### *AIM*

To investigate ChatGPT-4's accuracy and reproducibility in responding to patient questions related to SIBO.

### *METHODS*

A total of 27 patient questions related to SIBO were curated from professional societies, Facebook groups, and Reddit threads. Each question was entered into GPT-4 twice on separate days to examine reproducibility of accuracy on separate occasions. GPT-4 generated responses were independently evaluated for accuracy and reproducibility by two motility fellowship-trained gastroenterologists. A third senior fellowship-trained gastroenterologist resolved disagreements. Accuracy of responses were graded using the scale: (1) Comprehensive; (2) Correct but inadequate; (3) Some correct and some incorrect; or (4) Completely incorrect. Two responses were generated for every question to evaluate reproducibility in accuracy.

***RESULTS***
In evaluating GPT-4's effectiveness at answering SIBO-related questions, it provided responses with correct information to 18/27 (66.7%) of questions, with 16/27 (59.3%) of responses graded as comprehensive and 2/27 (7.4%) responses graded as correct but inadequate. The model provided responses with incorrect information to 9/27 (33.3%) of questions, with 4/27 (14.8%) of responses graded as completely incorrect and 5/27 (18.5%) of responses graded as mixed correct and incorrect data. Accuracy varied by question category, with questions related to "basic knowledge" achieving the highest proportion of comprehensive responses (90%) and no incorrect responses. On the other hand, the "treatment" related questions yielded the lowest proportion of comprehensive responses (33.3%) and highest percent of completely incorrect responses (33.3%). A total of 77.8% of questions yielded reproducible responses.

***CONCLUSION***
Though GPT-4 shows promise as a supplementary tool for SIBO-related patient education, the model requires further refinement and validation in subsequent iterations prior to its integration into patient care.

**Key Words:** Small intestinal bacterial overgrowth; Motility; Artificial intelligence; Chat-GPT; Large language models; Patient education



**Core Tip:** ChatGPT-4 demonstrates promise in enhancing patient understanding of basic concepts related to small intestinal bacterial overgrowth (SIBO). However, it exhibits limitations in accurately addressing questions about the diagnosis and treatment of SIBO, which are areas where up-to-date medical guidance is crucial. As such, artificial intelligence can be beneficial for general patient education but should not replace professional medical advice, especially for conditions with complex care protocols. Continuous refinement and updating of Chat-GPT's knowledge are essential for its safe and effective application in healthcare. Rigorous scrutiny of artificial intelligence-generated content is imperative to prevent the dissemination of potentially harmful misinformation.

**Citation:** Schlussel L, Samaan JS, Chan Y, Chang B, Yeo YH, Ng WH, Rezaie A. Evaluating the accuracy and reproducibility of ChatGPT-4 in answering patient questions related to small intestinal bacterial overgrowth. *Artif Intell Gastroenterol* 2024; 5(1): 90503
**URL:** https://www.wjgnet.com/2644-3236/full/v5/i1/90503.htm
**DOI:** https://dx.doi.org/10.35712/aig.v5.i1.90503

## INTRODUCTION

Small intestinal bacterial overgrowth (SIBO) is a medical condition characterized by an excessive amount of bacteria in the small intestine, which can lead to a variety of symptoms, including bloating, abdominal pain, diarrhea, and constipation[1]. The diagnosis and treatment of SIBO varies across institutions and by healthcare provider[2]. Though various tests exist, including glucose and lactulose breath tests and small intestine aspiration and culture, there is a lack of universal approach regarding how and when to utilize these tests, as well as how to interpret the results[3].

Due to the need for specialized tests, lack of dedicated International Classification of Diseases codes, and differences in the diagnostic methods across studies, it is challenging to estimate the prevalence of SIBO with studies showing rates ranging from 4% to 79%[2] and 38% to 84% in patients with IBS[4]. Importantly, SIBO has adverse effects on quality of life and may be associated with significant healthcare costs. Though the impact on quality of life for patients with SIBO has not been independently examined, one study showed that the presence of SIBO among patients with IBS was associated with more severe symptoms and led to a decreased quality of life[5]. Patients with IBS constitute a major proportion of patients who seek consultation in gastroenterology specialist clinic[6] and is associated with considerable healthcare resource use[7]. Given the high prevalence of SIBO among patients with IBS and its association with more severe symptoms, it's very likely that SIBO has a significant impact on patients and our healthcare system. Moreover, patients have limited access to motility specialists or physicians that are capable of managing SIBO, and may even encounter health care providers that question the legitimacy of SIBO as a medical condition[8].

The advent of artificial intelligence (AI) and natural language processing technologies has led to the development of large language models (LLMs), such as ChatGPT, which have the potential to revolutionize healthcare communication and patient education[9]. GPT-4, created by OpenAI, is able to produce easy to understand and conversational responses to inquiries by users based on their inquiries. It functions on the principle of predicting subsequent words in a sentence, much akin to an expert player in a game of 'guess the next word'[9]. There is a growing body of evidence demonstrating ChatGPT's ability to answer patient questions related to medical diseases such as cardiovascular disease, bariatric surgery and cirrhosis[10-12]. In a study comparing chatbot and physician responses, evaluators preferred chatbot answers 78.6% of the time[10]. The chatbot's responses were not only more comprehensive but also of higher quality and more empa-

thetic, with a 3.6 times higher prevalence of good or very good quality answers and a 9.8 times higher prevalence of empathetic or very empathetic responses than physicians[10]. Given the increasing trends of patients seeking healthcare related information from online sources, examining the strengths and limitations of LLMs as sources of information for patients is critical to ensuring safe, effective and responsible use of these models[13].

SIBO is a complex medical condition, with differing diagnostic and treatment approaches across institutions and healthcare providers as well as geographic variations in access to specialists. The gap in patient needs versus accessibility may lead individuals to seek information from alternative sources, such as the internet or ChatGPT. If proven safe and effective, emerging AI technologies like ChatGPT offer potential benefits in this space, providing accessible, easy to understand, and informed responses to patient inquiries, which may supplement or complement patient education provided by licensed healthcare professionals. In light of this, our study aimed to evaluate the accuracy of GPT-4 in providing accurate and reproducible responses to patient questions related to SIBO. This involved assessing the quality of information provided by the AI tool against evidence-based guidelines and expert opinions. Furthermore, our research will identify the limitations and potential risks associated with using GPT-4 as a supplementary tool for patient education and support, in order to inform the development of best practices for its implementation in the healthcare context.

## MATERIALS AND METHODS

### *Question curation*

A total of 38 patient questions related to SIBO were collected from professional societies and institutions as well as Facebook support groups ("SIBO lifestyle", "SIBO SOS Community") and the Reddit thread r/SIBO. Each question was screened to ensure it was directly related to SIBO. Questions that were not specific to SIBO or were outside the scope of typical patient concerns were excluded. Duplicate and similar questions were excluded to prevent redundancy and to ensure a broad coverage of topics. One question was removed after it was deemed incorrectly worded and containing incorrect information. The final set of 27 questions included in our study represents a diverse range of patient inquiries, covering aspects of basic knowledge, diagnosis, treatment, and other concerns related to SIBO.

### *ChatGPT*

ChatGPT is an AI LLM developed by OpenAI, based on the GPT (Generative Pre-trained Transformer) architecture. The model was designed to generate human-like text based on input, allowing the model to answer questions, engage in conversation, and perform various tasks. ChatGPT was trained on a large corpus of text from the internet, learning grammar, facts, and some reasoning abilities. It does not have a traditional "database" to retrieve information from; instead, the model generates text based on patterns and knowledge learned from the training data. However, it is essential to note the model's knowledge is limited to data up until September 2021, lacking awareness of more recent information. The latest iteration of the model, GPT-4, was released in March of 2023 and has shown promise across multiple domains of tasks[14].

### *Response generation*

GPT-4 was used on 4/23/23 and 4/24/23 to generate responses. Each question was entered as an individual prompt using the "New Chat" function. Each question was entered into GPT-4 twice on separate days to examine reproducibility of accuracy on separate occasions.

### *Response grading*

Reponses to questions were first independently graded for accuracy and reproducibility by two board certified, motility fellowship-trained, academic gastroenterologist reviewers actively practicing in a tertiary medical center. The following grading scale was used to grade the accuracy of each response similar to previous publications[11,12]: (1) Comprehensive (Grade 1): The response provides a complete and thorough answer as one would expect from a board-certified gastroenterologist. This grade implies that there is no additional relevant information that a specialist would deem necessary to include; (2) Correct but inadequate (Grade 2): The response is accurate but lacks certain critical details or depth that a board-certified gastroenterologist would consider important for a patient's understanding or management of SIBO; (3) Some correct and some incorrect (Grade 3): The response contains both correct and incorrect elements, indicating partial knowledge but with significant gaps or errors that require correction; and (4) Completely incorrect (Grade 4): The response does not provide accurate information related to the question asked and is considered misleading or wrong.

Reproducibility was graded based on the similarity in accuracy of the two responses per question generated by GPT-4. Any disagreement in reproducibility or accuracy grading was resolved by a third senior board-certified, motility fellowship trained gastroenterologist reviewer with greater than 10 years of experience in the field of gastrointestinal motility.

### *Statistical analysis*

Descriptive analysis is presented as counts and percentages. For statistical analysis purposes, questions were categorized into multiple subgroups: Basic knowledge, diagnosis, treatment, and others. All statistical analysis was performed in Excel version 2308.

## RESULTS

In total, 27 questions related to SIBO were inputted into GPT-4. The model provided 16/27 (59.3%) comprehensive, 2/27 (7.4%) correct but inadequate, 5/27 (18.5%) mixed with correct and incorrect data, and 4/27 (14.8%) completely incorrect responses. When examined by category, the model provided "comprehensive" responses to 90% of "basic knowledge questions", 60% of "diagnosis" questions, and 33.3% of "treatment" questions (Table 1). The model provided reproducible responses to 21/27 (77.8%) of questions (Table 2).

Most of the "completely incorrect" responses were noted to be in the "treatment" subcategory with 33.3% (3/9) of these responses rated as "completely incorrect". For example, when asked "What probiotic strain is recommended for constipation predominant SIBO?" GPT-4 stated that there is evidence that shows certain strains of probiotics helps constipation, which is not in line with current evidence and guidelines. Importantly, the model did recommend consulting with a health professional before starting new supplements. Questions, responses, and reviewer gradings are shown in Supplementary Table 1.

## DISCUSSION

SIBO is a common medical condition with variable approaches to management and diagnosis across institutions. The literature shows patients frequently pursue health-related information in lieu of their healthcare providers, with the internet emerging as a common source. Due to its user-friendly interface as well as its easy to understand and conversational responses, patients may utilize ChatGPT as a source of information regarding SIBO. In light of this, we examined ChatGPT's ability to accurately and reliably answer SIBO related questions. While the model provided comprehensive answers to 59.3% of questions, 14.8% of questions were graded as completely incorrect. Our findings show GPT-4's promising future in serving as an adjunct source of information for patient with SIBO but highlight its current limitations and need for further fine tuning, training and validation prior to incorporation into clinical care.

The model provided completely inaccurate responses to 4 (14.8%) questions and mixed correct and incorrect information to 5 (18.5%) questions, which is not in line with previous data which shows its proficiency in areas such as cirrhosis, congestive heart failure, and bariatric surgery[11,12,15]. For example, GPT-3.5 provided comprehensive responses to 86.8% of questions related to bariatric surgery and 83.2% of questions related to heart failure[11,15]. The reason for the difference in performance seen in our study may be related to the dataset used to train ChatGPT. There are well-established, thoroughly researched, and widely accepted guidelines governing the diagnosis and treatment of these conditions. Such robust guidelines offer a standardized framework, enabling ChatGPT to provide accurate and reliable responses. SIBO, however, presents a unique challenge due to its less definitive guidelines that often diverge across institutions and among physicians. Further compounding this issue is ChatGPT's knowledge constraints to information prior to 2021, restricting its ability to integrate the latest studies or consensus in the rapidly evolving field of SIBO and gut microbiome. This data limitation, paired with the inherent variability in SIBO management, showcases the system's vulnerabilities in areas where medical guidelines are either in flux or less established. Considering our analysis shows a considerable number of responses contained incorrect and potentially harmful information, this underscores the importance of exercising caution when utilizing AI-generated information in the context of patient education, particularly related to complex medical conditions like SIBO. Ongoing refinement and development of LLMs are imperative to mitigate the potential risks and enhance their potential role in patient education.

GPT-4 also showed a relatively low reproducibility, only delivering consistent accuracy of responses for 77.8% of questions. This again is in contrast with previous studies which found LLMs deliver high reproducibility of quality of responses[10-12]. Such reproducibility is critical for a tool intended to educate and inform, as consistent messaging is key in enhancing understanding, mitigating confusion and establishing trust among users.

Examining GPT-4's accuracy across different domains of patient questions allowed for a more granular analysis of its performance. In line with previous studies examining ChatGPT's knowledge in cirrhosis and hepatocellular carcinoma, bariatric surgery, and heart failure[11,12,15], we found GPT-4 provided comprehensive and accurate responses to the vast majority of basic knowledge questions. This suggests that AI has the potential to serve as a reliable resource of information for patients to enhance their basic understanding of their condition. Such an application aligns with a growing body of evidence pointing to the potential of AI in augmenting patient education[16]. However, our findings also underscore key limitations of this technology. Most notably, GPT-4's responses related to the diagnosis and treatment of SIBO contain a significant amount of inaccuracies. This finding is particularly concerning given that these areas often present the greatest challenges for patients in terms of understanding and self-management. Misinformation can lead to suboptimal patient decision-making and potential harm. It underlines the importance of caution when using AI for health-related advice and re-emphasizes the need for these tools to be used in tandem with professional medical guidance[9]. This suggests that while LLMs like GPT-4 in their current form may provide beneficial support for patients looking to enhance their general understanding of a condition, they are not yet equipped to offer reliable advice on more complex aspects of medical care. It is consistent with prior research noting the limitations of AI in understanding complex diseases and suggesting tailored, expert human intervention for such scenarios[17,18].

Beyond accuracy, comprehensiveness, and reproducibility, it's important to ensure LLMs produce materials that are easy to understand by patients of all health literacy levels. There is a growing body of literature showing LLMs are able to adjust the readability of outputs when prompted[19,20]. This ensures that access to information is democratized, and patients of all health literacy levels have personalized education materials. One study showed that GPT-4 was able to improve the readability of bariatric surgery patient education materials from 12$^{th}$ grade-college level to 6$^{th}$-9$^{th}$ grade[19].

**Table 1 Grading of responses generated by ChatGPT-4 to questions related to small intestinal bacterial overgrowth categorized by subgroup and overall**

| | % |
|---|---|
| Basic knowledge ($n$ = 10) | |
| Comprehensive | 90 |
| Correct but inadequate | 10 |
| Mixed with correct and incorrect data | 0 |
| Completely incorrect | 0 |
| Diagnosis ($n$ = 5) | |
| Comprehensive | 60 |
| Correct but inadequate | 0 |
| Mixed with correct and incorrect data | 40 |
| Completely incorrect | 0 |
| Treatment ($n$ = 9) | |
| Comprehensive | 33.3 |
| Correct but inadequate | 0 |
| Mixed with correct and incorrect data | 33.3 |
| Completely incorrect | 33.3 |
| Other ($n$ = 3) | |
| Comprehensive | 33.3 |
| Correct but inadequate | 33.3 |
| Mixed with correct and incorrect data | 0 |
| Completely incorrect | 33.3 |
| Overall ($n$ = 27) | |
| Comprehensive | 59.3 |
| Correct but inadequate | 7.4 |
| Mixed with correct and incorrect data | 18.5 |
| Completely incorrect | 14.8 |

**Table 2 Reproducibility of ChatGPT-4 responses overall and categorized by subgroup**

| | % |
|---|---|
| Overall ($n$ = 27) | 77.8 |
| Basic knowledge ($n$ = 10) | 90 |
| Diagnosis ($n$ = 5) | 80 |
| Treatment ($n$ = 9) | 77.8 |
| Other ($n$ = 3) | 33.3 |

Access to high quality patient education materials can also be impacted by patient language preference. Patients who prefer non-English languages have unique barriers to access to patient education materials. Some studies have shown the ability of LLMs in generating patient education materials in languages other than English with promising results[21-23]. Lastly, it's important to ensure outputs do not perpetuate known stereotypes and biases in medicine. There is a growing body of literature examining the presence of implicit bias in LLM outputs, with some studies showing LLMs may propagate racial and gender biases[24,25]. Future research should thoroughly investigate how LLMs can produce patient education materials that are not only accurate and of high quality but also accessible to patients from diverse backgrounds, with an emphasis on minimizing implicit bias and discrimination.

Limitations specific to the design of this study include the use of only two responses generated by GPT-4 to evaluate its reproducibility. While our findings provide initial insights, expanding the number of responses and questions in future research will be crucial to thoroughly assess consistency and reliability. Such expansions will help to substantiate the AI model's utility in patient education. Another limitation of this study is the use of the paid GPT-4 model over the free GPT-3.5, which was selected for its advanced linguistic capabilities and enhanced accuracy in medical contexts. While this choice aligns with our objective to evaluate the most current and sophisticated AI technology for patient education, it may affect the generalizability and accessibility of our findings. Future research could explore the trade-offs between cost and performance by comparing different AI models, including the cost-free GPT-3.5, to optimize the balance between accessibility and quality of information in AI-assisted patient care. Future studies would also benefit from exploring the differences in accuracy and reproducibility amongst different AI tools such as GPT-3.5, GPT-4, and Google Bard. For example, in a study comparing GPT-4 and Google Bard in their ability to diagnose and triage patients' ophthalmologic complaints, GPT-4 performed significantly better than Bard by generating more accurate triage suggestions, responses that experts were satisfied with for patient use, and lower potential harm rates[26]. Another study comparing GPT-3.5 and Bard in their ability to provide appropriate informational responses to patient questions regarding vascular surgery demonstrated that GPT-3.5 responses were more complete and more appropriate compared with Bard responses[27]. Similarly, GPT-3 exhibited greater accuracy and consistency over Google Bard, as well Google and Bing search engines, when addressing patient questions related to lung cancer[28]. These comparative evaluations underscore the evolving landscape of AI tools in healthcare and the importance of ongoing, meticulous analysis to harness their full potential for patient care.

Finally, we must consider other limitations of ChatGPT that pose a challenge for its future utilization in healthcare. OpenAI has not released specific details about the exact datasets used to train GPT-4. This raises concerns regarding the quality of data the model uses to respond to questions, especially when discussing healthcare related topics. The literature in healthcare is rapidly evolving and requires staying up to date with the literature to ensure good practice of medicine. ChatGPT's lack of continuous updates limits its generalized applicability in patient care. Another constraint of GPT-4 and LLMs in general is the "hallucination effect," where the model produces outputs that seem plausible and believable but are incorrect, misleading, or entirely fabricated[29,30]. This is a significant limitation that should be considered when implementing such AI tools in the healthcare setting. Our study design also has its limitations. Responses from ChatGPT were graded based on expert opinion which is subjective and prone to bias. Notably, this is a limitation across the majority of literature examining the clinical knowledge of ChatGPT, given expert opinion guided by the literature and guidelines is currently the gold standard in the practice of medicine. Our study utilized a sample of 27 patient questions, which is not inclusive of all possible patient questions pertaining to SIBO. We performed a systematic approach when curating questions to reduce the risk of selection bias. Furthermore, questions were not removed after the generation of responses from ChatGPT.

## CONCLUSION

Our study underscores the potential future value of large language models, like GPT-4, in patient education related to SIBO, especially in providing basic knowledge. However, we highlight the limitations of GPT-4 in its current form due to a significant number of its responses containing inaccurate or out of date information and low reproducibility in accuracy of its responses. While AI may supplement traditional patient education methods in the future, it is not a substitute for professional medical advice. Continued evaluation and development of these technologies are crucial to harness their potential while minimizing potential harm. This iterative process will be key to the future integration of AI into healthcare systems, with the ultimate aim of improving patient understanding, engagement, and outcomes. Our research underscores the need for rigorous scrutiny and cautious application when relying on AI technologies for diseases with complex and less standardized diagnosis and treatments. Given the prevalence and impact of SIBO on patients and the healthcare system, the need for accurate, accessible patient education remains critical. Our research serves as a valuable step in identifying the challenges and opportunities for integrating AI tools in this capacity.

## FOOTNOTES

**Author contributions:** Rezaie A was the guarantor, participated in the acquisition, analysis, and interpretation of the data, and revised the article for critically important intellectual content; Schlussel L drafted the initial manuscript and participated in the acquisition, analysis, and interpretation of the data; Samaan J designed the study and revised the article for critically important intellectual content; Chan Y and Chang B participated in the acquisition, analysis, and interpretation of the data, and revised the article for critically important intellectual content; Yeo YH revised the article for critically important intellectual content; Ng WH participated in the acquisition, analysis, and interpretation of the data.

**Institutional review board statement:** Our study did not require IRB approval, given our research does not involve human subjects.

**Informed consent statement:** Our research does not involve human subjects. Therefore, no signed informed consent or documents were obtained.

**Conflict-of-interest statement:** All the authors declare that they have no conflict of interest.

**Data sharing statement:** No additional data are available.

**STROBE statement:** The authors have read the STROBE Statement – checklist of items, and the manuscript was prepared and revised according to the STROBE Statement – checklist of items.



**ORCID number:** Lauren Schlussel 0009-0000-9008-4460; Yin Chan 0000-0003-2741-8258; Yee Hui Yeo 0000-0002-2703-5954; Ali Rezaie 0000-0002-0106-372X.

**Corresponding Author's Membership in Professional Societies:** American College of Gastroenterology, No. 68989.

**S-Editor:** Liu JH
**L-Editor:** A
**P-Editor:** Zhao YQ

## REFERENCES

1 **Sachdev AH**, Pimentel M. Gastrointestinal bacterial overgrowth: pathogenesis and clinical significance. *Ther Adv Chronic Dis* 2013; **4**: 223-231 [PMID: 23997926 DOI: 10.1177/2040622313496126]
2 **Rao SSC**, Bhagatwala J. Small Intestinal Bacterial Overgrowth: Clinical Features and Therapeutic Management. *Clin Transl Gastroenterol* 2019; **10**: e00078 [PMID: 31584459 DOI: 10.14309/ctg.0000000000000078]
3 **Rezaie A**, Pimentel M, Rao SS. How to Test and Treat Small Intestinal Bacterial Overgrowth: an Evidence-Based Approach. *Curr Gastroenterol Rep* 2016; **18**: 8 [PMID: 26780631 DOI: 10.1007/s11894-015-0482-9]
4 **Posserud I**, Stotzer PO, Björnsson ES, Abrahamsson H, Simrén M. Small intestinal bacterial overgrowth in patients with irritable bowel syndrome. *Gut* 2007; **56**: 802-808 [PMID: 17148502 DOI: 10.1136/gut.2006.108712]
5 **Chuah KH**, Hian WX, Lim SZ, Beh KH, Mahadeva S. Impact of small intestinal bacterial overgrowth on symptoms and quality of life in irritable bowel syndrome. *J Dig Dis* 2023; **24**: 194-202 [PMID: 37200005 DOI: 10.1111/1751-2980.13189]
6 **Chuah KH**, Cheong SY, Lim SZ, Mahadeva S. Functional dyspepsia leads to more healthcare utilization in secondary care compared with other functional gastrointestinal disorders. *J Dig Dis* 2022; **23**: 111-117 [PMID: 35050547 DOI: 10.1111/1751-2980.13082]
7 **Canavan C**, West J, Card T. Review article: the economic impact of the irritable bowel syndrome. *Aliment Pharmacol Ther* 2014; **40**: 1023-1034 [PMID: 25199904 DOI: 10.1111/apt.12938]
8 **Ruscio M**. Is SIBO A Real Condition? *Altern Ther Health Med* 2019; **25**: 30-38 [PMID: 31550680]
9 **Gilson A**, Safranek CW, Huang T, Socrates V, Chi L, Taylor RA, Chartash D. How Does ChatGPT Perform on the United States Medical Licensing Examination (USMLE)? The Implications of Large Language Models for Medical Education and Knowledge Assessment. *JMIR Med Educ* 2023; **9**: e45312 [PMID: 36753318 DOI: 10.2196/45312]
10 **Ayers JW**, Poliak A, Dredze M, Leas EC, Zhu Z, Kelley JB, Faix DJ, Goodman AM, Longhurst CA, Hogarth M, Smith DM. Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum. *JAMA Intern Med* 2023; **183**: 589-596 [PMID: 37115527 DOI: 10.1001/jamainternmed.2023.1838]
11 **Samaan JS**, Yeo YH, Rajeev N, Hawley L, Abel S, Ng WH, Srinivasan N, Park J, Burch M, Watson R, Liran O, Samakar K. Assessing the Accuracy of Responses by the Language Model ChatGPT to Questions Regarding Bariatric Surgery. *Obes Surg* 2023; **33**: 1790-1796 [PMID: 37106269 DOI: 10.1007/s11695-023-06603-5]
12 **Yeo YH**, Samaan JS, Ng WH, Ting PS, Trivedi H, Vipani A, Ayoub W, Yang JD, Liran O, Spiegel B, Kuo A. Assessing the performance of ChatGPT in answering questions regarding cirrhosis and hepatocellular carcinoma. *Clin Mol Hepatol* 2023; **29**: 721-732 [PMID: 36946005 DOI: 10.3350/cmh.2023.0089]
13 **Cima RR**, Anderson KJ, Larson DW, Dozois EJ, Hassan I, Sandborn WJ, Loftus EV, Pemberton JH. Internet use by patients in an inflammatory bowel disease specialty clinic. *Inflamm Bowel Dis* 2007; **13**: 1266-1270 [PMID: 17567877 DOI: 10.1002/ibd.20198]
14 **OpenAI**. GPT-4 Technical Report. 2023 [DOI: 10.48550/arXiv.2303.08774]
15 **King RC**, Samaan JS, Yeo YH, Mody B, Lombardo DM, Ghashghaei R. Appropriateness of ChatGPT in answering heart failure related questions. 2023. Available from: https://www.medrxiv.org/content/10.1101/2023.07.07.23292385v1
16 **Ayre J**, Mac O, McCaffery K, McKay BR, Liu M, Shi Y, Rezwan A, Dunn AG. New Frontiers in Health Literacy: Using ChatGPT to Simplify Health Information for People in the Community. *J Gen Intern Med* 2024; **39**: 573-577 [PMID: 37940756 DOI: 10.1007/s11606-023-08469-w]
17 **Alkaissi H**, McFarlane SI. Artificial Hallucinations in ChatGPT: Implications in Scientific Writing. *Cureus* 2023; **15**: e35179 [PMID: 36811129 DOI: 10.7759/cureus.35179]
18 **Dave T**, Athaluri SA, Singh S. ChatGPT in medicine: an overview of its applications, advantages, limitations, future prospects, and ethical considerations. *Front Artif Intell* 2023; **6**: 1169595 [PMID: 37215063 DOI: 10.3389/frai.2023.1169595]
19 **Srinivasan N**, Samaan JS, Rajeev ND, Kanu MU, Yeo YH, Samakar K. Large language models and bariatric surgery patient education: a comparative readability analysis of GPT-3.5, GPT-4, Bard, and online institutional resources. *Surg Endosc* 2024 [PMID: 38472531 DOI: 10.1007/s00464-024-10720-2]
20 **Rouhi AD**, Ghanem YK, Yolchieva L, Saleh Z, Joshi H, Moccia MC, Suarez-Pierre A, Han JJ. Can Artificial Intelligence Improve the Readability of Patient Education Materials on Aortic Stenosis? A Pilot Study. *Cardiol Ther* 2024; **13**: 137-147 [PMID: 38194058 DOI: 10.1007/s40119-023-00347-0]

21 **Yeo YH**, Samaan JS, Ng WH, Ma X, Ting P, Kwak M, Panduro A, Lizaola-Mayo B, Trivedi H, Vipani A, Ayoub W, Yang JD, Liran O, Spiegel B, Kuo A. GPT-4 outperforms ChatGPT in answering non-English questions related to cirrhosis. 2023. Available from: https://www.medrxiv.org/content/10.1101/2023.05.04.23289482v1

22 **Samaan JS**, Yeo YH, Ng WH, Ting PS, Trivedi H, Vipani A, Yang JD, Liran O, Spiegel B, Kuo A, Ayoub WS. ChatGPT's ability to comprehend and answer cirrhosis related questions in Arabic. *Arab J Gastroenterol* 2023; **24**: 145-148 [PMID: 37673708 DOI: 10.1016/j.ajg.2023.08.001]

23 **Wang H**, Wu W, Dou Z, He L, Yang L. Performance and exploration of ChatGPT in medical examination, records and education in Chinese: Pave the way for medical AI. *Int J Med Inform* 2023; **177**: 105173 [PMID: 37549499 DOI: 10.1016/j.ijmedinf.2023.105173]

24 **Omiye JA**, Lester JC, Spichak S, Rotemberg V, Daneshjou R. Large language models propagate race-based medicine. *NPJ Digit Med* 2023; **6**: 195 [PMID: 37864012 DOI: 10.1038/s41746-023-00939-z]

25 **Kaplan DM**, Palitsky R, Arconada Alvarez SJ, Pozzo NS, Greenleaf MN, Atkinson CA, Lam WA. What's in a Name? Experimental Evidence of Gender Bias in Recommendation Letters Generated by ChatGPT. *J Med Internet Res* 2024; **26**: e51837 [PMID: 38441945 DOI: 10.2196/51837]

26 **Zandi R**, Fahey JD, Drakopoulos M, Bryan JM, Dong S, Bryar PJ, Bidwell AE, Bowen RC, Lavine JA, Mirza RG. Exploring Diagnostic Precision and Triage Proficiency: A Comparative Study of GPT-4 and Bard in Addressing Common Ophthalmic Complaints. *Bioengineering (Basel)* 2024; **11** [PMID: 38391606 DOI: 10.3390/bioengineering11020120]

27 **Chervonski E**, Harish KB, Rockman CB, Sadek M, Teter KA, Jacobowitz GR, Berland TL, Lohr J, Moore C, Maldonado TS. Generative artificial intelligence chatbots may provide appropriate informational responses to common vascular surgery questions by patients. *Vascular* 2024; 17085381241240550 [PMID: 38500300 DOI: 10.1177/17085381241240550]

28 **Rahsepar AA**, Tavakoli N, Kim GHJ, Hassani C, Abtin F, Bedayat A. How AI Responds to Common Lung Cancer Questions: ChatGPT vs Google Bard. *Radiology* 2023; **307**: e230922 [PMID: 37310252 DOI: 10.1148/radiol.230922]

29 **Shen Y**, Heacock L, Elias J, Hentel KD, Reig B, Shih G, Moy L. ChatGPT and Other Large Language Models Are Double-edged Swords. *Radiology* 2023; **307**: e230163 [PMID: 36700838 DOI: 10.1148/radiol.230163]

30 **Xiao Y**, Wang WY. On Hallucination and Predictive Uncertainty in Conditional Language Generation. In: Merlo P, Tiedemann J, Tsarfaty R, eds. Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume. Association for Computational Linguistics; 2021: 2734-2744 [DOI: 10.18653/v1/2021.eacl-main.236]

Published by **Baishideng Publishing Group Inc**
7041 Koll Center Parkway, Suite 160, Pleasanton, CA 94566, USA
**Telephone:** +1-925-3991568
**E-mail:** office@baishideng.com
**Help Desk:** https://www.f6publishing.com/helpdesk
https://www.wjgnet.com