W J G

World Journal of Gastroenterology

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

World J Gastroenterol 2024 August 7; 30(29): 3538-3540

DOI: 10.3748/wjg.v30.i29.3538

ISSN 1007-9327 (print) ISSN 2219-2840 (online)

LETTER TO THE EDITOR

Evaluating the role of large language models in inflammatory bowel disease patient information

Eun Jeong Gong, Chang Seok Bang

Specialty type: Gastroenterology and hepatology

Provenance and peer review: Invited article; Externally peer reviewed.

Peer-review model: Single blind

Peer-review report's classification
Scientific Quality: Grade B
Novelty: Grade A
Creativity or Innovation: Grade B
Scientific Significance: Grade B

P-Reviewer: Dai YC

Received: May 27, 2024
Revised: July 15, 2024
Accepted: July 22, 2024
Published online: August 7, 2024
Processing time: 62 Days and 22 Hours

Eun Jeong Gong, Chang Seok Bang, Department of Internal Medicine, Hallym University College of Medicine, Chuncheon 24253, Gangwon-do, South Korea

Corresponding author: Chang Seok Bang, MD, PhD, Associate Professor, Doctor, Department of Internal Medicine, Hallym University College of Medicine, Sakju-ro 77, Chuncheon 24253, Gangwon-do, South Korea. csbang@hallym.ac.kr

Abstract

This letter evaluates the article by Gravina *et al* on ChatGPT's potential in providing medical information for inflammatory bowel disease patients. While promising, it highlights the need for advanced techniques like reasoning + action and retrieval-augmented generation to improve accuracy and reliability. Emphasizing that simple question and answer testing is insufficient, it calls for more nuanced evaluation methods to truly gauge large language models' capabilities in clinical applications.

Key Words: Crohn's disease; Ulcerative colitis; Inflammatory bowel disease; Chat generative pre-trained transformer; Large language model; Artificial intelligence

Core Tip: This commentary evaluates the article by Gravina *et al* on ChatGPT's potential in providing medical information for inflammatory bowel disease patients. While promising, it highlights the need for advanced techniques like reasoning + action and retrieval-augmented generation to improve accuracy, emphasizing that simple question-and-answer testing is insufficient for evaluating large language models' true capabilities.

Citation: Gong EJ, Bang CS. Evaluating the role of large language models in inflammatory bowel disease patient information. *World J Gastroenterol* 2024; 30(29): 3538-3540
URL: https://www.wjgnet.com/1007-9327/full/v30/i29/3538.htm
DOI: https://dx.doi.org/10.3748/wjg.v30.i29.3538

TO THE EDITOR

We are writing to express out thoughts on the recently published article by Gravina *et al*[1]. Gravina *et al*[1] assessed the capability of large language models (LLMs) like ChatGPT to provide plausible medical information to patients with inflammatory bowel disease (IBD). Despite identifying several limitations, the authors concluded that there is significant potential in using LLMs for this purpose[1].

One of the key insights from the article is the potential for ChatGPT to offer immediate and accessible information to patients. The authors correctly note that this could be particularly beneficial in providing preliminary guidance and answering common queries that patients may have about their condition. This aligns with the increasing trend of patients seeking health information online before consulting their healthcare providers.

However, the study also underscores significant limitations, such as the potential for outdated or inaccurate information. Given that medical knowledge is continuously evolving, it is crucial for artificial intelligence (AI) tools like ChatGPT to have mechanisms for regular updates to ensure the information provided is current and evidence-based[1]. This is especially important for chronic conditions like IBD, where treatment guidelines and best practices frequently change.

A pertinent question arises: Can LLMs truly perform inference? Current AI-based agents utilizing LLMs operate by either generating answers directly or referring to external tools if the LLM itself cannot provide an answer. These agents determine the necessary information, redefine the questions, call appropriate tools to extract information, analyze the extracted data, and iterate this process as needed to reach a final answer. This pattern, known as reasoning + action, closely mimics human problem-solving by iteratively refining questions and seeking relevant tools rather than merely retrieving similar past solutions[2].

The effectiveness of such an approach often hinges on prompt engineering. Enhanced prompt engineering can significantly improve the accuracy of LLM-generated answers by aligning queries more closely with the model's trained data and inference capabilities. Therefore, evaluating LLMs based on selected questions often reflects their proficiency in leveraging search tools to produce desired answers. Advanced prompt engineering techniques can potentially yield more accurate responses, indicating that simple question-and-answer testing might not fully capture an LLM's capabilities[3].

Moreover, the retrieval-augmented generation (RAG) technique enhances traditional LLMs by enabling real-time retrieval of external data not included in the training dataset, thus generating answers that integrate the latest information. This approach helps prevent hallucination and allows the model to utilize a broader knowledge base. However, standardized performance evaluation of these advanced techniques remains challenging due to the limited benchmarks available, making it difficult to assess using only a few representative questions[4].

Another important point raised by the authors is the issue of contextual understanding and empathy, which AI currently lacks. The physician patient relationship is built on trust and understanding, and while AI can provide factual information, it cannot replace the nuanced, empathetic communication that healthcare providers offer. This aspect is particularly vital for managing chronic diseases that significantly impact patients' quality of life[1].

The authors' recommendation for further refinement and alignment of AI outputs with reliable medical databases is essential. Such improvements could enhance the accuracy and reliability of AI-generated medical information, making it a more robust tool for both patients and healthcare providers.

Despite these challenges, there is no doubt that LLMs, equipped with sophisticated learning datasets and RAG capabilities, hold promise for clinical application. However, evaluating their potential solely based on simple question-answer accuracy is inadequate. It is essential to consider the advanced techniques and iterative processes that significantly enhance the precision and reliability of LLM-generated medical information.

In conclusion, the article by Gravina *et al*[1] provides valuable insights into the current capabilities and limitations of AI in gastroenterology. While promising, further refinement and a more nuanced evaluation approach are crucial for realizing the full potential of AI in healthcare. Continued research and development, combined with rigorous validation against established medical standards, will be essential.

FOOTNOTES

Author contributions: Gong EJ and Bang CS contributed to conceptualization, methodology, investigation and wrote the original draft; Bang CS reviewed and edited the draft, and contributed to supervision; All authors have read and agreed to the published version of the manuscript.

Conflict-of-interest statement: The authors declare that they have no conflict of interest.

Country of origin: South Korea

ORCID number: Eun Jeong Gong 0000-0003-3996-3472; Chang Seok Bang 0000-0003-4908-5431.

S-Editor: Fan M

L-Editor: A
P-Editor: Yu HG

REFERENCES

1 **Gravina AG**, Pellegrino R, Cipullo M, Palladino G, Imperio G, Ventura A, Auletta S, Ciamarra P, Federico A. May ChatGPT be a tool producing medical information for common inflammatory bowel disease patients' questions? An evidence-controlled analysis. *World J Gastroenterol* 2024; **30**: 17-33 [PMID: 38293321 DOI: 10.3748/wjg.v30.i1.17]

2 **Verma M**, Bhambri S, Kambhampati S. On the Brittle Foundations of ReAct Prompting for Agentic Large Language Models. 2024 Preprint. Available from: arXiv 2405. 13966 [DOI: 10.48550/arXiv.2405.13966]

3 **Kim HJ**, Gong EJ, Bang CS. Application of Machine Learning Based on Structured Medical Data in Gastroenterology. *Biomimetics (Basel)* 2023; **8**: 512 [PMID: 37999153 DOI: 10.3390/biomimetics8070512]

4 **Guinet G**, Omidvar-Tehrani B, Deoras A, Callot L. Automated Evaluation of Retrieval-Augmented Language Models with Task-Specific Exam Generation. 2024 Preprint. Available from: arXiv 2405. 13622 [DOI: 10.48550/arXiv.2405.13622]

Baishideng®

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