

# Revolutionizing gastrointestinal endoscopy: the emerging role of large language models

Eun Jeong Gong<sup>1,2,3</sup>, Chang Seok Bang<sup>1,2,3</sup>

<sup>1</sup>Department of Internal Medicine, Hallym University College of Medicine, Chuncheon; <sup>2</sup>Institute for Liver and Digestive Diseases, Hallym University, Chuncheon; <sup>3</sup>Institute of New Frontier Research, Hallym University College of Medicine, Chuncheon, Korea

As healthcare continues to evolve with technological advancements, the integration of artificial intelligence (AI) into clinical practice has shown promising potential for enhancing patient care and operational efficiency.<sup>1-3</sup> Large language models (LLMs), as a subset of AI technologies, are at the forefront of this revolution, offering capabilities that extend far beyond simple data processing. These models possess the unique ability to understand, generate, and interact with human language on an unprecedented scale, thereby opening new avenues for enhancing the clinical practices across various specialties, including gastroenterology.

Gastrointestinal (GI) endoscopy, a cornerstone procedure for the diagnosis and treatment of digestive tract disorders, is used to integrate advanced technologies. Endoscopic procedures rely on the expertise of specialists in interpreting complex visual data and performing precise interventions. This presents a unique set of challenges, such as the variability in diagnostic accuracy and a labor-intensive reporting and documentation process. LLMs, with their vast data processing capabilities, promise to address these challenges by enhancing the diagnostic accuracy, automating report generation, enabling clinical reasoning,

and improving educational tools.<sup>4</sup> This editorial aims to explore the emerging role of LLMs in the field of GI endoscopy and provide future directions of this symbiotic relationship between AI technology and GI endoscopy.

## POTENTIAL ROLE OF LLMS IN GI ENDOSCOPY

The advent of LLMs heralds a new era in GI endoscopy marked by improved diagnostic accuracy, streamlined documentation, and enhanced educational and patient engagement strategies. By analyzing endoscopic images with unparalleled precision and automatically generating reports, LLMs introduce a level of analysis and efficiency that was previously unattainable, which could reduce diagnostic errors and administrative burdens. Their ability to quickly assess, interpret, and synthesize large volumes of medical data can transform the diagnostic process, providing support for complex medical queries, including rare or obscure conditions, and keeping medical professionals abreast of the latest research. LLMs can also be used for clinical reasoning.

In addition to the diagnostics and report generation, LLMs show significant potential in the education domain. They can create interactive training materials, personalize patient education, and provide emotional support.<sup>5,6</sup> Emotional support is another unexpected benefit of LLMs, language vision models, and foundation models with multimodal functions. These capabilities can improve patient experience, enhance satisfaction, and foster a deeper understanding of medical conditions and treatments, and are expected to be the next-generation mainstream AI technology in clinical practice.

Received: February 17, 2024    Revised: March 10, 2024

Accepted: March 11, 2024

Correspondence: Chang Seok Bang

Department of Internal Medicine, Hallym University Chuncheon Sacred Heart Hospital, Hallym University College of Medicine, 77 Sakju-ro, Chuncheon 24253, Korea

E-mail: csbang@hallym.ac.kr

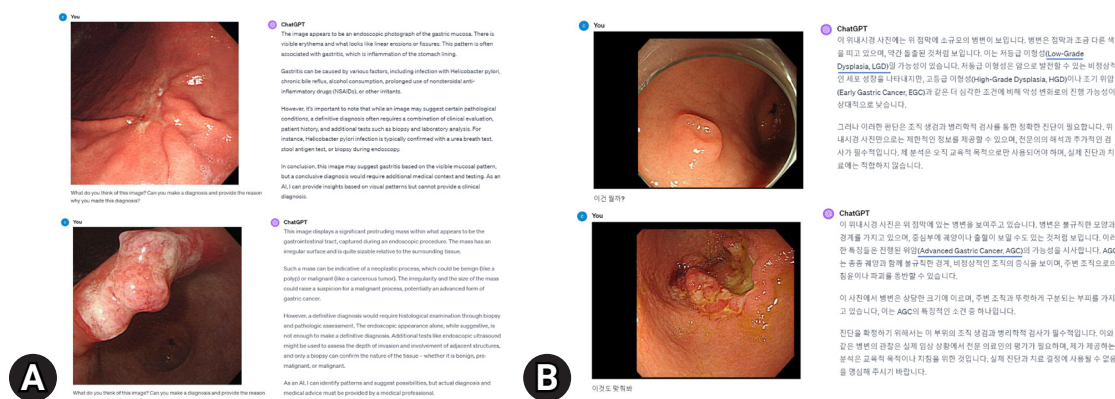
© This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<https://creativecommons.org/licenses/by-nc/4.0/>) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

## BENEFITS AND LIMITATIONS

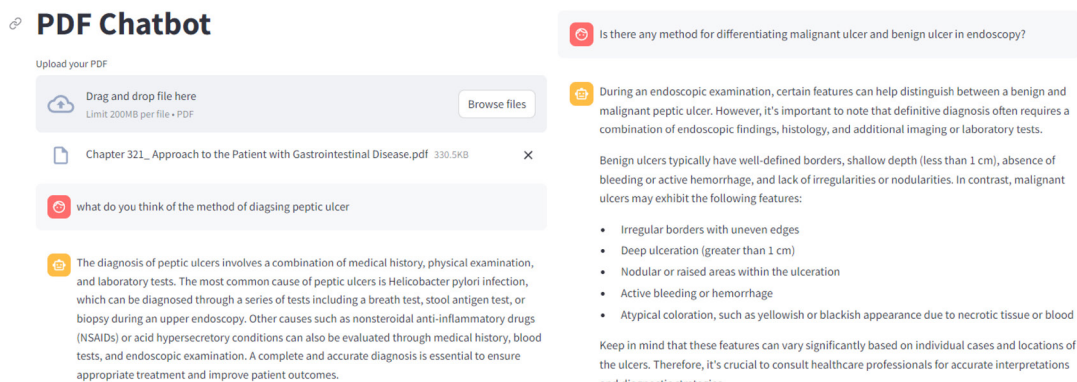
The integration of LLMs into GI endoscopy promises to bring numerous benefits, including but not limited to, enhanced diagnostic accuracy, efficiency in clinical operations, and improved patient engagement (Fig. 1).<sup>4</sup> The ability of the LLMs to serve as a dynamic source of knowledge for both medical staff and patients facilitates better communication, supports research, and contributes to quality improvement in medical practice

(Figs. 2–4).<sup>7</sup> For instance, the use of LLMs to analyze electronic medical records to identify patients for specific interventions or to understand the quality metrics, such as adenoma detection rates through pathology reports exemplifies their potential to revolutionize clinical practices.<sup>8</sup>

However, the implementation of LLMs is challenging. Data privacy, biases in training data, necessity for interdisciplinary collaboration, and the need for human oversight remain significant hurdles. The technical and cultural barriers of integrating



**Fig. 1.** (A) An example of zero-shot learning in a large language model. The upper panel shows a still-cut image of a poorly differentiated stomach adenocarcinoma. The zero-shot learning results (answers) were incorrect. The lower panel shows a still-cut image of submucosal invasion of early gastric cancer (adenocarcinoma). The zero-shot learning results (answers) were incorrect. This analysis was performed using ChatGPT 4 in January 2024. (B) An example of one-shot learning in a large language model. We trained the model using one representative image each of low-grade dysplasia, high-grade dysplasia, early gastric cancer, and advanced gastric cancer. The upper panel shows a still-cut image of low-grade dysplasia of the stomach, where the one-shot learning result was correct. The lower panel presents a still-cut image of advanced gastric cancer, with correct one-shot learning results. This analysis was performed using ChatGPT 4 in January 2024.



**Fig. 2.** A still cut image of a local PDF chat application developed using the Mistral 7-B large language model, Langchain, Ollama, and Streamlit (<https://github.com/SonicWarrior1/pdfchat>). The language model incorporates the concept of retrieval-augmented generation, which allows it to produce responses in the context of specific documents. It demonstrated appropriate question and answer capabilities after analyzing Chapter 321 of the 21st edition of Harrison's Principles of Internal Medicine.



**Fig. 3.** A schematic view of a questionnaire system designed to facilitate communication between medical staff and patients (provided by Alexis Reality Co., Ltd.).

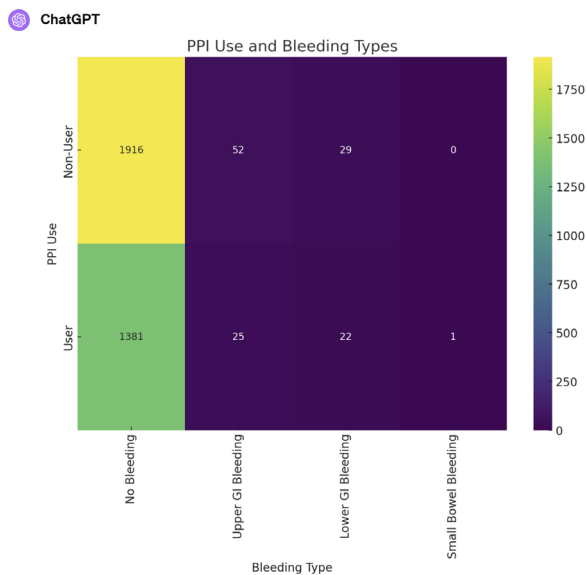


Table 1: Clinical Characteristics of the Enrolled Patients

Variable	Non-PPI Users (N=1998)	PPI Users (N=1429)
Age (years)	71.87 ± 12.68	73.60 ± 11.07
Sex (M/F)	73/79	79/71
Antithrombotics	3.22 ± 1.74	3.85 ± 1.61
Bleeding	0.06 ± 0.29	0.05 ± 0.28
DM (Yes/No)	619/1379	498/928
Heart Failure (Yes/No)	421/1577	455/974
A.Fib (Yes/No)	1135/861	941/486

Note: Age is presented as mean ± standard deviation. For categorical variables, the table presents the number of patients in each category.

**Fig. 4.** An example of the analysis function in a large language model. ChatGPT 4 was accessed for this purpose in February 2024.

these technologies into clinical practice should be addressed along with the ethical considerations of AI use in healthcare. Recent studies have highlighted the propensity of LLMs to amplify societal biases and overrepresent stereotypes, raising concerns about the equitable application of AI technologies.<sup>9</sup> Rather than relying solely on the recommendations produced by the model, it is essential that the model be connected to an independent, verifiable source of bias-free knowledge via retrieval-augmented generation.<sup>9,10</sup>

Another major consideration was the likelihood of hallucinations. LLM is a general-purpose model. Prompt engineering, in which explicit instructions meant to exploit the optimal capabilities of LLMs are incorporated in addition to the question within the LLM input, can significantly improve the LLM performance for specific tasks. However, hallucinations may become common if humans simply ask questions without providing specific instructions. Fine-tuning or retrieval-augmented generation may improve the goal directedness of LLMs.<sup>4</sup>

## FUTURE DIRECTIONS

The future of LLMs in GI endoscopy is poised to be at the intersection of AI technology, interdisciplinary collaboration, and ethical governance. As the field advances, it will be crucial to focus on patient-centric innovations and leverage LLMs to address global health disparities. The development of transformer-based language vision models and their potential to cover current convolutional neural network (CNN)-based approaches in GI endoscopy illustrates the ongoing evolution of AI technologies. Medical practice is a multimodal task that includes history taking, visual diagnosis, data interpretation, and clinical reasoning. Notably, LLM-based foundation models that have multimodal functions are the next-generation mainstream AI models in clinical practice.<sup>4</sup>

LLMs are one of the forms of generative models; however, generative models other than text-generation (generative adversarial network, diffusion, variational autoencoder, language vision models, etc.) are also evolving, and the creative features in these models are already being integrated into LLMs. The emergence of foundation models with multimodal functions underscores the shift towards more integrated and comprehensive AI tools in clinical practice, promising a future in which the capabilities of AI models are fully harnessed to enhance diagnostic and therapeutic outcomes in the field of GI endoscopy. We are currently developing new LLM models with significant-

ly larger parameters and optimizations. Although they have the potential to enhance our practice, it is essential to handle them carefully to avoid potential harm.<sup>4</sup>

## CONCLUSION

The integration of LLMs into GI endoscopy represents a frontier of healthcare innovation with the potential to significantly enhance the diagnostic accuracy, operational efficiency, and patient care. However, this journey is contingent on overcoming the challenges of data privacy, ensuring the quality of the data used for AI training, and fostering interdisciplinary collaboration.

## Conflicts of Interest

Chang Seok Bang is currently serving as a KSGE Publication Committee member; however, he was not involved in peer reviewer selection, evaluation, or the decision process in this study. The other author has no potential conflicts of interest.

## Funding

None.

## Author Contributions

Conceptualization: CSB; Investigation: EJG; Resources: EJG; Writing-original draft: all authors; Writing-review & editing: all authors.

## ORCID

Eun Jeong Gong	<a href="https://orcid.org/0000-0003-3996-3472">https://orcid.org/0000-0003-3996-3472</a>
Chang Seok Bang	<a href="https://orcid.org/0000-0003-4908-5431">https://orcid.org/0000-0003-4908-5431</a>

## REFERENCES

1. Bang CS. Artificial intelligence in the analysis of upper gastrointestinal disorders. *Korean J Helicobacter Up Gastrointest Res* 2021;21:300–310.
2. Gong EJ, Bang CS, Lee JJ, et al. Clinical decision support system for all stages of gastric carcinogenesis in real-time endoscopy: model establishment and validation study. *J Med Internet Res* 2023;25:e50448.
3. Gong EJ, Bang CS, Lee JJ, et al. Deep learning-based clinical decision support system for gastric neoplasms in real-time endoscopy: development and validation study. *Endoscopy* 2023;55:701–708.
4. Kim HJ, Gong EJ, Bang CS. Application of machine learning based on structured medical data in gastroenterology. *Biomimetics (Basel)* 2023;8:512.
5. Buzzaccarini G, Degliuomini RS, Borin M, et al. The promise and pitfalls of AI-generated anatomical images: evaluating midjourney for aesthetic surgery applications. *Aesthetic Plast Surg* 2024;48:1874–1883.
6. Chin H, Song H, Baek G, et al. The potential of chatbots for emotional support and promoting mental well-being in different cultures: mixed methods study. *J Med Internet Res* 2023;25:e51712.
7. Ge J, Sun S, Owens J, et al. Development of a liver disease-specific large language model chat Interface using retrieval augmented generation. *Hepatology* 2024;80:1158–1168.
8. Savage T, Wang J, Shieh L. A large language model screening tool to target patients for best practice alerts: development and validation. *JMIR Med Inform* 2023;11:e49886.
9. Zack T, Lehman E, Suzgun M, et al. Assessing the potential of GPT-4 to perpetuate racial and gender biases in health care: a model evaluation study. *Lancet Digit Health* 2024;6:e12–e22.
10. Hastings J. Preventing harm from non-conscious bias in medical generative AI. *Lancet Digit Health* 2024;6:e2–e3.