Infection and Drug Resistance

Dovepress
Taylor & Francis Group

Open Access Full Text Article

ORIGINAL RESEARCH

Performance of Large Language Models in Chinese Language Medical Counseling on *Helicobacter pylori*

Mingjun Zhang*, Shiming Zhou*, Shulin Zhang, Ting Yi, Bo Jiang, Xuan Jiang

Department of Gastroenterology, Beijing Tsinghua Changgung Hospital, School of Clinical Medicine, Tsinghua Medicine, Tsinghua University, Beijing, People's Republic of China

*These authors contributed equally to this work

Correspondence: Xuan Jiang; Bo Jiang, Email jxa01998@btch.edu.cn; drjiang@163.com

Background: *H. pylori* infection is a worldwide health issue, fueling rising demand for medical counseling. LLMs have the potential to serve in medical counseling. However, their performance remains unclear.

Objective: This study aimed to evaluate the effectiveness of LLMs in providing *H. pylori* related medical counseling in a Chinese context.

Methods: 20 *H. pylori*-related questions were collected, covering four domains: definition and symptoms, diagnosis, treatment, and prevention. Each question was asked thrice in Chinese to each LLM. We assessed the responses across five dimensions (accuracy, relevance, completeness, clarity, and reliability).

Results: 1. In the first batch of tests, the overall performance distribution was 33.3% good, 66.1% medium, and 0.6% poor, respectively. No significant differences were observed among the three LLMs (p=0.158). Good performance was observed with 47.8% in accuracy, 53.9% in relevance, 68.3% in completeness, 36.7% in clarity, and 36.1% in reliability. No significant differences were observed in accuracy, relevance, completeness, or clarity. Reliability differed significantly (p<0.001), with Ernie Bot achieving the best performance. 2. The second test batch yielded performance rates of 70.6% good, 29.4% medium, and 0% poor, with a significant difference among the three LLMs (p=0.018). Doubao attained the best performance, surpassing other models in relevance and clarity. 3. The newly assessed AI batch showed markedly superior overall performance to the counterpart evaluated more than a year prior.

Conclusion: This study is the first to evaluate the effectiveness of various LLMs in *H. pylori*-related medical counseling in a real-world setting. The study showed that while LLMs generally performed acceptably in terms of accuracy, relevance, and completeness, their clarity and reliability were less satisfactory. Ernie Bot, developed by Chinese company, outperformed ChatGPT in certain aspects of medical counseling in Chinese. With the guidance of professionals, LLMs can serve as potential aids for medical counseling.

Keywords: artificial intelligence, large language model, medical counseling, *Helicobacter pylori*

Introduction

Helicobacter pylori (*H. pylori*) infection is a global health concern, with its prevalence varying widely across countries, ranging from 20% to 50% in high-income countries to more than 80% in some low-income countries.[1] *H. pylori* infection causes chronic gastritis and increases the risk of peptic ulcers, gastric cancer, and mucosa-associated lymphoid tissue lymphoma.[2] In 2020, gastric cancer accounted for more than 1 million new cases and approximately 770 000 deaths, with China alone accounting for approximately half of these new cases.[3] Concerns regarding *H. pylori* infection are rising among the Chinese population, leading to an increased demand for medical counseling. The high prevalence of *H. pylori* infection poses a substantial challenge to public health, largely due to a lack of health awareness and education. Health education not only affects individual health outcomes, but also has a broader socioeconomic impact.

Medical counseling plays an important role in facilitating communication between patients and health care professionals, enabling them to better understand their health status and make informed health management decisions. The shortage of professionals leads to inefficiency in health counseling and disease screening. Meeting the rapid growth in people's health management needs is an urgent problem that needs to be solved.

Received: 6 August 2025
Accepted: 15 December 2025
Published: 20 December 2025

Large Language Models (LLMs) are artificial intelligence models that are trained on extensive textual data to generate human-like responses. LLMs have been utilized in various medical applications, from answering health inquiry to generating clinical reports.[4,5] With continuous improvements, LLMs show great promise in enhancing healthcare delivery.

The Superbench Large Model Comprehensive Capability Evaluation Report, jointly released by the Basic Model Research Center of Tsinghua University and Zhongguancun Laboratory, provided an open, dynamic, scientific, and authoritative evaluation of large models based on five benchmarks (ExtremeGLUE, NaturalCodeBench, AlignBench, AgentBench, and SafetyBench). The report affirmed the ability of LLMs to understand inputs and generate outputs stably, and indicated that LLMs developed by Chinese companies outperform their foreign counterparts in a Chinese setting. In practice, the performance of LLMs has been found to be language-dependent.[6] In addition, caution is warranted regarding the phenomenon known as "artificial intelligence (AI) hallucinations".[7]

Given the potential of LLMs in healthcare communication, this study aimed to evaluate the effectiveness of *H. pylori*-related medical counseling in China.

Methods

Data Source and Study Design

The subjects of this study were AI systems (non-human participants), the clinical questions used were fictional or fully de-identified cases (not involving identifiable patient information), the research activities had no privacy risks and were of the lowest risk level. Therefore, this study met the ethical exemption criteria.

The first batch of data generation for this study was conducted in August 2024. Based on the March 2024 edition of the Superbench Large Model Comprehensive Capability Evaluation Report, we selected three well performed LLMs: ChatGPT 3.5 turbo (OpenAI), Kimi (Moonshot AI), and Ernie Bot 3.5 (Baidu, Inc), with the latter two developed by Chinese companies and designed for the Chinese language.

The second batch of data generation for this study was completed in November 2025. Three AI models were incorporated in this phase, including two developed by Chinese companies (Doubao from ByteDance; DeepSeek-V3 from Hangzhou Deep Search Artificial Intelligence Basic Technology Research Co., Ltd) and one foreign AI model (Gemini 2.5 Pro from Google LLC).

Three board-certified physicians participated in this study. Based on an overview of the clinical guidelines for *H. pylori* infection and personal experiences during face-to-face, telephone, and doctor-patient news portal interactions, we selected 20 questions covering four domains: definition and symptoms, diagnosis, treatment, and prevention.

Response Generation and Grading

We collected 20 questions about which patients were most concerned, covering aspects such as symptoms, diagnosis, treatment, and prevention. These 20 questions were collected from real-world conversations between patients and doctors in the outpatient department (detailed in supplementary document 1). Twenty questions were compiled into a set and presented three times for each LLM. The answers generated from each set of questions were labeled as iterations A, B, and C. We analyzed the results of three iterations to evaluate the consistency of the LLM responses. All browsing data, including cookies, were cleared after each iteration to avoid bias from the correlation interference.

All evaluations were guided by the Sixth Chinese National Consensus Report on the Management of *Helicobacter pylori* infection (treatment excluded)[8] and Management of *Helicobacter pylori* infection: the Maastricht VI/Florence consensus report.[9] Responses were recorded and blindly assigned to three physicians, who assessed them on the following five dimensions using a 5-point Likert scale:

Accuracy

This dimension evaluates whether the answer is correct, without factual errors. A score of 5 indicated complete accuracy without any errors, and a score of 1 indicated that the answer contained either critical or outright errors.

Relevance

This dimension evaluates whether the answer directly addresses the question, and does not deviate from the topic. A score of 5 indicated that the response was completely related to the question, and a score of 1 indicated that the answer was irrelevant to the question or completely off the topic.

Completeness

This dimension evaluates whether the answer is comprehensive and covers all key points of the question. A score of 5 indicated a thorough answer that covered all necessary information, and a score of 1 indicated that the answer was incomplete and missing key information.

Clarity

This dimension evaluates how clearly the answer is expressed, and its ease of understanding. A score of 5 indicated that the response was clear, well-articulated, and easy for readers to understand, and a score of 1 indicated that the response was confusing or difficult to understand.

Reliability

This dimension evaluates whether the answer is credible based on the quality of the information or logical reasoning provided. A score of 5 reflects a highly reliable response grounded in trustworthy information or sound reasoning. A score of 1 indicated low confidence in the response because of insufficient support or poor logic.

Overall evaluation: The overall evaluation was comprehensively derived based on these scores. An average score of four or higher was classified as good, an average score between three and four (not included) was classified as medium, and an average score below three was classified as poor.

All outputs were scored independently by three physicians, and the average score was used for model evaluation. The intra-class correlation coefficient (ICC) values and 95% confidence intervals of the five-dimensional scores of the three physicians were 0.732[0.672, 0.785], 0.681[0.613, 0.741], 0.698[0.632, 0.756], 0.655[0.583, 0.720], and 0.612[0.534, 0.684], respectively, which were considered moderately consistent. In addition, an assessment by one of the physicians was used to evaluate the stability of the LLMs by determining whether there were differences among the responses to the same question.

Data Analysis

SPSS 27.0 statistical software was employed for the analysis. Categorical data were described using frequencies and percentages and tested using the chi-square test or Fisher's exact probability method. Standard deviation was used to evaluate the stability of the three LLMs. Statistical significance was set at $p<0.05$.

Results

Analysis of the Results from the First Batch of Data

Overall Evaluation

The overall distribution of good, medium, and poor performances was 33.3% (60 instances), 66.1% (119 instances), and 0.6% (1 instance), respectively. The performances of the three LLMs are shown in Figure 1, with no significant differences ($p=0.158$).

Sub-Index Evaluation

Accuracy

The performances of the three LLMs in terms of accuracy are shown in Figure 2, with no significant differences ($p=0.521$). The overall distribution of good, medium, and poor performance was 47.8% (86 instances), 51.7% (93 instances), and 0.6% (1 instance), respectively.

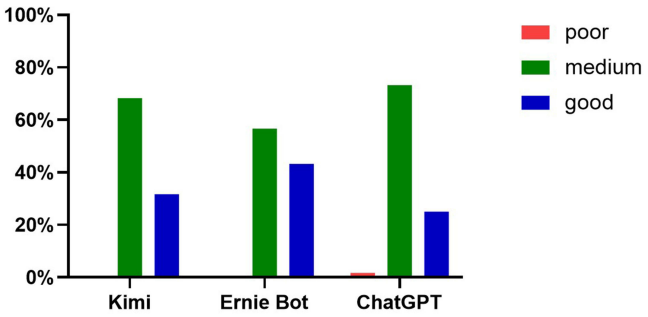

Figure 1 The overall performances of the three LLMs.

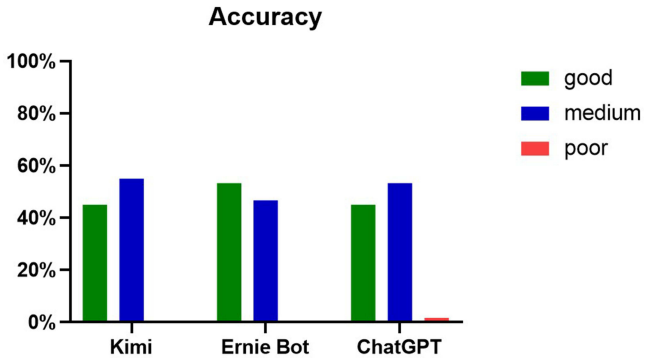

Figure 2 The performances of the three LLMs in terms of accuracy.

Relevance

The respective performances of the three LLMs in terms of relevance are shown in Figure 3, with no significant differences (p=0.382). The overall distribution of good, medium, and poor performances was 53.9% (97 instances), 46.1% (83 instances), and 0% (0 instance), respectively.

Completeness

The respective performances of the three LLMs in terms of completeness are shown in Figure 4, with no significant differences (p=0.150). The overall distribution of good, medium, and poor performances was 68.3% (123 instances), 31.1% (56 instances), and 0.6% (1 instance), respectively.

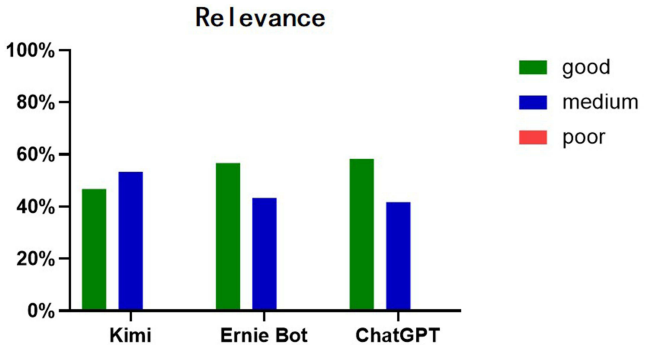

Figure 3 The performances of the three LLMs in terms of relevance.

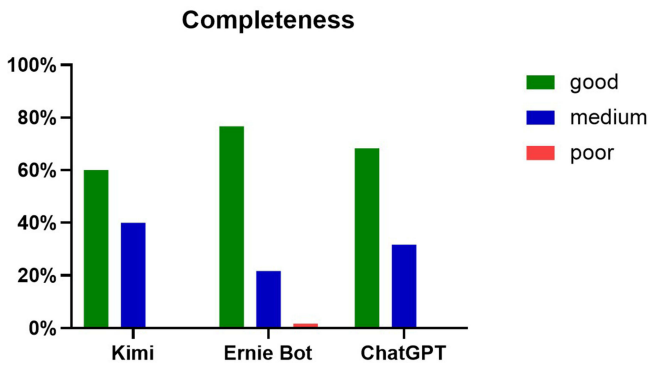

Figure 4 The performances of the three LLMs in terms of completeness.

Clarity

The respective performances of the three LLMs, in terms of clarity, are shown in Figure 5, with no significant differences (p=0.385). The overall distribution of good, medium, and poor performances was 36.7% (66 instances), 68.2% (111 instances), and 1.7% (3 instances), respectively.

Reliability

The respective performances of the three LLMs in terms of reliability are shown in Figure 6, with significant differences (p<0.001). ERNIE Bot performed the best, with a good performance rate of 55%. The overall distribution of good, medium, and poor performances was 36.1% (65 instances), 56.7% (102 instances), and 7.2% (13 instances), respectively.

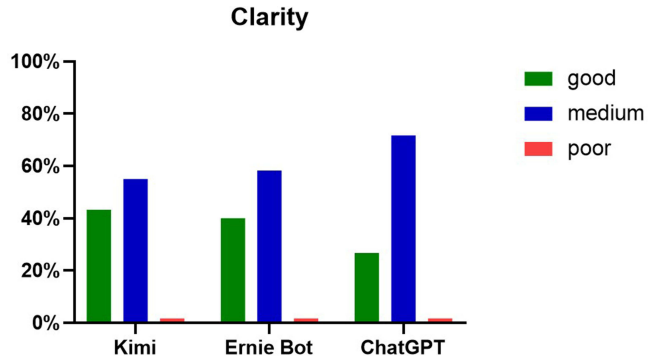

Figure 5 The performances of the three LLMs, in terms of clarity.

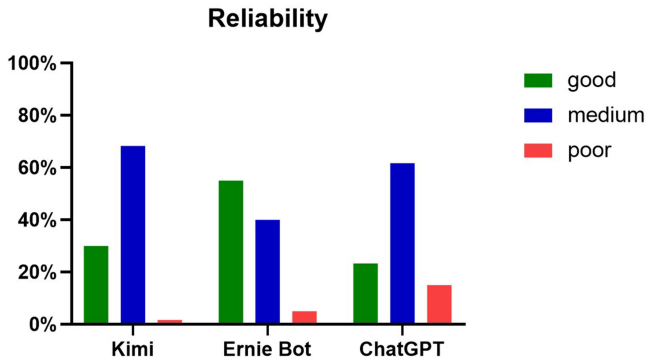

Figure 6 The performances of the three LLMs in terms of reliability.

https://doi.org/10.2147/IDR.S553523

Performance of the three LLMs in four different domains (definition and symptoms, diagnosis, treatment, and prevention).

The performance on the responses of the definition and symptoms part is shown in Table 1, with no statistical difference among the three LLMs (p=0.095).

The performance on responses of the diagnosis part is shown in Table 2. There is a statistical difference (p=0.0.021), and ERNIE Bot performs the best.

The performance on responses of the treatment part is shown in Table 3, with no statistical difference among the three LLMs (p=0.160).

The performance on responses of the prevention part is shown in Table 4, with no statistical difference among the three LLMs (p=0.219).

Analysis of the Results from the Second Batch of Data

Overall Evaluation

The overall distribution of good, medium, and poor performances was 70.6% (127 instances), 29.4% (53 instances), and 0%, respectively. There was a statistically significant difference among the three LLMs (χ^2=9.076, p=0.018), with Doubao achieving the best performance.

Table 1 The Performance on the Responses of the Definition and Symptoms Part

	Good (score≥4)	Medium (3≤score<4)	Poor (score<3)
Kimi	7 (58.3%)	5 (41.7%)	0 (0%)
Ernie Bot	10 (83.3%)	2 (16.7%)	0 (0%)
ChatGPT	5 (41.7%)	7 (58.3%)	0 (0%)

Table 2 The Performance on Responses of the Diagnosis Part

	Good (score≥4)	Medium (3≤score<4)	Poor (score<3)
Kimi	4 (26.7%)	11 (73.3%)	0 (0%)
Ernie Bot	11 (73.3%)	4 (26.7%)	0 (0%)
ChatGPT	5 (33.3%)	10 (66.7%)	0 (0%)

Table 3 The Performance on Responses of the Treatment Part

	Good (score≥4)	Medium (3≤score<4)	Poor (score<3)
Kimi	2 (11.1%)	16 (88.9%)	0 (0%)
Ernie Bot	3 (16.7%)	15 (83.3%)	0 (0%)
ChatGPT	0 (0%)	17 (94.4%)	1 (5.6%)

Table 4 The Performance on Responses of the Prevention Part

	Good (score≥4)	Medium (3≤score<4)	Poor (score<3)
Kimi	6 (40%)	9 (60%)	0 (0%)
Ernie Bot	2 (13.3%)	13 (86.7%)	0 (0%)
ChatGPT	5 (33.3%)	10 (66.7%)	0 (0%)

Sub-Index Evaluation

The performances of the three LLMs in terms of accuracy showed no significant differences (χ^2=8.261, p=0.018). The overall distribution of good, medium, and poor performance was 61.1% (110 instances), 37.8% (68 instances), and 1.1% (2 instances), respectively.

The performances of the three LLMs in terms of relevance showed significant differences (χ^2=40.774, p=0.000), with Doubao performed the best. The overall distribution of good, medium, and poor performances was 66.7% (120 instances), 32.8% (59 instances), and 0.56% (1 instance), respectively.

In terms of completeness, all three LLMs were rated as good in their performance on every single question.

The performances of the three LLMs in terms of clarity showed significant differences (χ^2=29.371, p=0.000), with Doubao performed the best. The overall distribution of good, medium, and poor performances was 73.3% (132 instances), 26.1% (47 instances), and 0.56% (1 instance), respectively.

The performances of the three LLMs in terms of clarity showed significant differences (χ^2=14.329, p=0.006), with Gemini performed the best. The overall distribution of good, medium, and poor performances was 57.2% (103 instances), 41.1% (74 instances), and 1.67% (3 instance), respectively.

Output Stability Evaluation

Based on an assessment by one of the physicians, the stability of the three responses from each LLM for every question was assessed. The average standard deviations of the three outputs were 0.376 (Kimi), 0.262 (Ernie Bot), 0.360 (ChatGPT), 0.412 (DeepSeek), 0.330 (Doubao), and 0.399 (Gemini), and the stability of the Ernie Bot was the best.

Comparison Between the Two Datasets

The overall performance scores of the two groups of LLMs showed a statistically significant difference (χ^2=51.920, p=0.000). The newly assessed AI batch exhibited markedly superior overall performance compared with the counterpart evaluated over a year prior.

Discussion

This study is the first to evaluate the effectiveness of various LLMs in *H. pylori*-related medical counseling in a real-world setting. All interactions were conducted in Chinese, not English, and required a higher level of linguistic comprehension. This study showed that LLMs performed commendably during *H. pylori*-related medical counseling. The prevalence of *H. pylori* infections in China is high. With the guidance of professionals, LLMs can serve as potential aids for medical counseling.

The AI models we tested were all publicly available, easily searchable by the general public, and had not undergone pre-training or fine-tuning. Overall, the responses of LLMs to *H. pylori*-related medical counseling in China were acceptable and provided valuable medical guidance to the general public. However, their performance was not entirely satisfactory, which may be linked to the language factors. At present, most LLMs are designed based on English, and they receive more training in English than in other languages. Studies have demonstrated that LLMs tend to perform better in English comprehension and output than Chinese ones in medical practice.[6,10] Therefore, we speculate that LLMs designed based on the Chinese language may be more advantageous for Chinese medical counseling. This study demonstrated that LLMs developed by Chinese companies outperformed their counterparts from English-speaking countries in several respects.

LLMs can potentially reduce costs while maintaining patient satisfaction and medical efficiency, and have broad application prospects in the medical field.[11] However, there are also potential risks and challenges, such as ethical and social impacts including bias, privacy, misinformation transmission, AI hallucinations.[12]

In this study, LLMs received high scores in terms of completeness. LLMs tend to output all the relevant content and provide more comprehensive answers. However, excessive information can lead to a decline in logical coherence and clarity. As shown in this study, LLMs performed unsatisfactorily in terms of clarity, indicating that AI's linguistic competence of AI is still not paired with that of humans. Although LLMs excel at formal linguistic competence

(knowledge of linguistic rules and patterns), their performance on functional linguistic competence (knowledge of linguistic rules and patterns) tasks remains spotty.[13]

It is noteworthy that LLMs performed poorly in terms of reliability. Owing to the complexity of AI language models, it is difficult to ensure interpretability and transparency. Online health information can lead to the dissemination of false information, possibly endangering individual health.[14] The tendency of LLMs to hallucinate in varying degrees in both content and references is disturbing.[15–17] During the scoring process, we also found instances in which LLM outputs were "imagined" without literature or data support, known as AI hallucinations. For instance, when answering the question of whether *Helicobacter pylori* recurs after treatment, ChatGPT mentioned in one of its responses that "the recurrence rate of *Helicobacter pylori* after treatment is usually below 10%", but this does not have definite theoretical support. For another example, ChatGPT mentioned that "it is necessary to wait for 4–6 weeks after the treatment when conducting blood tests to confirm whether the eradication was successful", but this claim lacks a theoretical basis, as antibodies are not suitable for monitoring *H. pylori* status following its eradication. Identifying and restricting hallucinations remain a significant concern for professionals. Reasonable policies and regulations are necessary to ensure safe and effective application of LLMs in medical practice. Healthcare professionals can play an important role in integrating AI into medicine,[7] emphasizing the need for professional oversight in the application of LLMs in the medical field.

The datasets generated or analyzed during this study are available from the corresponding author upon reasonable request.

This study identified a notable phenomenon: the overall scores of the evaluation data completed in 2025 were consistently higher than those obtained in 2024. Given the extremely rapid pace of AI iteration and updates, this result is not unexpected. With the continuous advancement of AI, its performance will also keep improving.

Despite these encouraging results, this study had several limitations. In this study, only three doctors participated in the scoring process, which inevitably led to subjective bias. This study evaluated only three AI models and the sample size was relatively small. Additionally, the study evaluated 20 questions that may not fully represent the spectrum of *H. pylori*-related medical counseling queries encountered in clinical practice. In future, we will expand the sample size and collect more questions from real-world patient consultations and electronic health records. This will help enhance the general applicability of this study. More experts should be invited to participate in the scoring process and increase the types of diseases under study to explore the feasibility of using AI in medical consulting. In addition, user studies can be conducted among actual healthcare professionals or patients to evaluate the credibility and usefulness of the AI models.

In conclusion, LLMs demonstrate great potential for medical counseling, although professional review and further evaluation are required. At present, these general AI models cannot be directly integrated into medical workflows; however, they can currently be used for patient education, preliminary screening, and assisting in clinical decision-making for primary care physicians.

Acknowledgment

This paper was presented at JMIR Publications as a preprint: Hyperlink with DOI (10.2196/preprints.68692), but was not accepted.

Disclosure

The authors report no conflicts of interest in this work.

References

1. Yang L, Kartsonaki C, Yao P, et al. The relative and attributable risks of cardia and non-cardia gastric cancer associated with Helicobacter pylori infection in China: a case-cohort study. *Lancet Public Health*. 2021;6(12):e888–e896. doi:10.1016/S2468-2667(21)00164-X
2. FitzGerald R, Smith SM. An overview of Helicobacter pylori infection. *Methods Mol Biol*. 2021;2283:1–14.
3. Sung H, Ferlay J, Siegel RL, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin*. 2021;71(3):209–249. doi:10.3322/caac.21660
4. Omiye JA, Gui H, Rezaei SJ, et al. Large language models in medicine: the potentials and pitfalls: a narrative review. *Ann Intern Med*. 2024;177(2):210–220. doi:10.7326/M23-2772
5. Liu PR, Lu L, Zhang JY, et al. Application of artificial intelligence in medicine: an overview. *Curr Med Sci*. 2021;41(6):1105–1115. doi:10.1007/s11596-021-2474-3

6. Kong QZ, Ju KP, Wan M, et al. Comparative analysis of large language models in medical counseling: a focus on Helicobacter pylori infection. *Helicobacter*. 2024;29(1):e13055. doi:10.1111/hel.13055
7. Hatem R, Simmons B, Thornton JE. A call to address AI "Hallucinations" and how healthcare professionals can mitigate their risks. *Cureus*. 2023;15(9):e44720. doi:10.7759/cureus.44720
8. Helicobacter pylori Study Group, Chinese Society of Gastroenterology, Chinese Medical Association. Sixth Chinese national consensus report on the management of Helicobacter pylori infection(treatment excluded)[J]. *Chin J Dig*. 2022;42(5):289–303.
9. European Helicobacter and Microbiota Study group; Malfertheiner P, Megraud F, Rokkas T, et al. Management of Helicobacter pylori infection: the maastricht VI/Florence consensus report. *Gut*. 2022;gutjnl–2022–327745. doi:10.1136/gutjnl-2022-327745
10. Wang H, Wu W, Dou Z, et al. 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. doi:10.1016/j.ijmedinf.2023.105173
11. Thirunavukarasu AJ, Ting DSJ, Elangovan K, et al. Large language models in medicine. *Nat Med*. 2023;29(8):1930–1940. doi:10.1038/s41591-023-02448-8
12. Zhui L, Fenghe L, Xuehu W, et al. Ethical considerations and fundamental principles of large language models in medical education: viewpoint. *J Med Internet Res*. 2024;26:e60083. doi:10.2196/60083
13. Mahowald K, Ivanova AA, Blank IA, et al. Dissociating language and thought in large language models. *Trends Cognit Sci*. 2024;28(6):517–540. doi:10.1016/j.tics.2024.01.011
14. Daraz L, Morrow AS, Ponce OJ, et al. Readability of online health information: a meta-narrative systematic review. *Am J Med Qual*. 2018;33(5):487–492. doi:10.1177/1062860617751639
15. Eysenbach G. The role of ChatGPT, generative language models, and artificial intelligence in medical education: a conversation with ChatGPT and a call for papers. *JMIR Med Educ*. 2023;9:e46885. doi:10.2196/46885
16. Aljamaan F, Temsah MH, Altamimi I, et al. Reference hallucination score for medical artificial intelligence chatbots: development and usability study. *JMIR Med Inform*. 2024;12:e54345. doi:10.2196/54345
17. Athaluri SA, Manthena SV, Kesapragada VSRKM, et al. Exploring the boundaries of reality: investigating the phenomenon of artificial intelligence hallucination in scientific writing through ChatGPT references. *Cureus*. 2023;15(4):e37432. doi:10.7759/cureus.37432