The performance of large language models in Helicobacter pylori related medical counseling within a Chinese-language context: A comparative analysis

Mingjun Zhang, Xuan Jiang

Submitted to: JMIR Formative Research
on: November 12, 2024

Table of Contents

The performance of large language models in Helicobacter pylori related medical counseling within a Chinese-language context: A comparative analysis

Mingjun Zhang[1] MD; Xuan Jiang[1]

[1]Beijing Tsinghua Chang Gung Hospital Beijing CN

Corresponding Author:
Xuan Jiang
Beijing Tsinghua Chang Gung Hospital
168 Litang Road?Changping District
Beijing
CN

Abstract

Background: Helicobacter pylori (H. pylori) infection is a global health issue, leading to a growing demand for medical counseling. Large Language Models (LLMs) have the potential to serve as valuable auxiliary tools for medical counseling. However, their performance in terms of accuracy, relevance, completeness, clarity, and reliability in providing H. pylori related medical counseling within a Chinese-language context remains unclear.

Objective: This study aimed to evaluate the effectiveness of three LLMs: ChatGPT 3.5 turbo?OpenAI?, Kimi?Moonshot AI?, and Ernie Bot 3.5(Baidu, Inc), in providing H. pylori related medical counseling in a Chinese context.

Methods: A total of 20 H. pylori related questions were included covering the following domains: definition and symptoms, diagnosis, treatment and prevention. Each question was being asked three times in Chinese to each LLM. An evaluation team of 3 physicians assessed the responses using a Likert scale across 5 dimensions: accuracy, relevance, completeness, clarity, and reliability, and came up with an overall assessment.

Results: A total of 20 H. pylori related questions were included covering the following domains: definition and symptoms, diagnosis, treatment and prevention. Each question was being asked three times in Chinese to each LLM. An evaluation team of 3 physicians assessed the responses using a Likert scale across 5 dimensions: accuracy, relevance, completeness, clarity, and reliability, and came up with an overall assessment.

Conclusions: This study is the first to evaluate the effectiveness of various LLMs in H. pylori related medical counseling in a real-world setting. All the interactions were conducted in Chinese, not English, demanding a higher level of linguistic comprehension. The study showed that while the LLMs generally performed acceptably in accuracy, relevance and completeness, their clarity and reliability were less satisfactory. Kimi and Ernie Bot, both developed by Chinese companies, outperformed ChatGPT in some aspects of medical counseling in Chinese language. With the guidance of professionals, LLMs can serve as potential aids for medical counseling.

(JMIR Preprints 12/11/2024:68692)
DOI: https://doi.org/10.2196/preprints.68692

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**

Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

The performance of large language models in Helicobacter pylori related medical counseling within a Chinese-language context: A comparative analysis

Mingjun Zhang, Shiming Zhou, Ting Yi, Xuan Jiang
Mingjun Zhang and Shiming Zhou contributed equally to the article
corresponding author: Xuan Jiang, jxa01998@btch.edu.cn
Beijing Tsinghua Changgung Hospital, School of Clinical Medicine, Tsinghua University

Abstract

Background

Helicobacter pylori (H. pylori) infection is a global health issue, leading to a growing demand for medical counseling. Large Language Models (LLMs) have the potential to serve as valuable auxiliary tools for medical counseling. However, their performance in terms of accuracy, relevance, completeness, clarity, and reliability in providing H. pylori related medical counseling within a Chinese-language context remains unclear.

Objective

This study aimed to evaluate the effectiveness of three LLMs: ChatGPT 3.5 turbo（OpenAI）, Kimi（Moonshot AI）, and Ernie Bot 3.5(Baidu, Inc), in providing H. pylori related medical counseling in a Chinese context.

Methods

A total of 20 H. pylori related questions were included covering the following domains: definition and symptoms, diagnosis, treatment and prevention. Each question was being asked three times in Chinese to each LLM. An evaluation team of 3 physicians assessed the responses using a Likert scale across 5 dimensions: accuracy, relevance, completeness, clarity, and reliability, and came up with an overall assessment.

Results

1. The overall distribution of good, medium and poor performance across all models was 33.3% (60 instances), 66.1% (119 instances) and 0.6% (1 instance), respectively. There was no significant difference between the three LLMs (P=0.12).

2.Good performance was observed in 47.8% (86 instances) for accuracy，53.9% (97 instances) for relevance, 68.9% (124 instances) for completeness，36.7% (66 instances) for clarity and 36.1% (65 instances) for clarity. While the three LLMs showed no significant differences in accuracy, relevance, completeness and clarity. A significant difference was noted in reliability (P＜0.001),with Ernie Bot performing the best.

3. Ernie Bot outperformed the others in the diagnosis domain, while all three LLMs showed comparable performance in definition and symptom, treatment and prevention.

4. Kimi and Ernie Bot demonstrated stable outputs, whereas ChatGPT exhibited less consistent performance.

Conclusion

This study is the first to evaluate the effectiveness of various LLMs in H. pylori related medical counseling in a real-world setting. All the interactions were conducted in Chinese, not English, demanding a higher level of linguistic comprehension. The study showed that while the LLMs generally performed acceptably in accuracy, relevance and completeness, their clarity and reliability

were less satisfactory. Kimi and Ernie Bot, both developed by Chinese companies, outperformed ChatGPT in some aspects of medical counseling in Chinese language. 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

H. pylori infection is a global health concern, with 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 ulcer, 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 about H. pylori infection are rising among the Chinese population, leading to increased demands 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 healthcare professionals, enabling patients to better understand their health status and make informed health management decisions. The shortage of professionals leads to the inefficiency of health counseling and disease screening. How to meet the rapid growth of people's health management needs is an urgent problem to be solved.

LLMs are artificial intelligence models trained on extensive amounts of 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 generating 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 also 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 performance on H. pylori related medical counseling in Chinese language.

Methods

Data source and study design

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.

Three board-certified physicians participated in the study. Based on an overview of the clinical guidelines for H. pylori 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

The 20 questions were compiled into a set and then presented three times to 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 LLM responses. All browsing data, including cookies, was cleared after each iteration to avoid bias from 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：

1. Accuracy: This dimension evaluates whether the answer is completely correct without factual errors. A score of 5 indicates that the answer is complete accuracy without any errors; A score of 1 indicates that the answer contains either critical errors or outright errors.

2. Relevance: This dimension evaluates whether the answer directly addresses the question and does not deviate from the topic. A score of 5 indicates that the response is completely related to the question; A score of 1 indicates that the answer that is irrelevant to the question or completely off topic.

3. Completeness: This dimension evaluates whether the answer is comprehensive and covers all the key points of the question. A score of 5 indicates a thorough answer that covers all necessary information; A score of 1 means that the answer is incomplete and missing key information.

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

5. 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 indicates low confidence, in the response due to insufficient support, or poor logic.

Overall evaluation: The overall evaluation is derived comprehensively on the above scores. An average score of 4 or above is classified as good, an average score between 3 and 4 (not included) is classified as medium, and an average score below 3 is classified as poor.

All outputs were scored independently by 3 physicians with the average score

used for model evaluation. In addition, the assessment of one of the physicians were used to evaluate the stability of the LLMs by determining whether there are differences among the three responses to the same question.

The study complied with the ethical standards outlined in the Helsinki Declaration and complied with national regulations in their respective fields. Since the study did not involve the use of human or animal data, it did not require approval from an ethics committee.

Data Analysis

SPSS 27.0 statistical software was employed for analysis. Categorical data were described using frequency and percentage, tested by chi-square test or Fisher's exact probability method. A repeated measure ANOVA was used to evaluate the stability of the three LLMs, respectively. $P<0.05$ indicated a statistically significant difference.

Results

1.Overall evaluation: The average scores of the three LLMs for 60 answers to 20 questions were shown in Table 1. The overall distribution of good, medium and poor performance was 33.3% (60 instances), 66.1% (119 instances) and 0.6% (1 instance), respectively. There were no significant differences among the three LLMs (P=0.12).

Table 1□

	good(score≥4)	medium □3≤score □4□	poor□score□3□
Kimi	19□31.7%□	41□68.3%□	0□0%□
Ernie Bot	26□43.3%□	34□56.7%□	0□0%□
ChatGPT	15□25%□	44□73.3%□	1□1.7%□

2.Sub-index evaluation□

2.1 Accuracy

The accuracy scores of the three LLMs were shown in Table 2, with no significant differences (P=0.422). The overall distribution of good, medium and poor performance was 47.8% (86 instances), 51.1% (92 instances) and 1.1% (2 instances), respectively.

Table 2

	good(score≥4)	medium □3≤score □4□	poor□score□3□
Kimi	27□45%□	33□55%□	0□0%□
Ernie Bot	32□53.3%□	28□46.7%□	0□0%□
ChatGPT	27□45%□	31□51.7%□	2□3.3%□

2.2 Correlation

The correlation scores of the three LLMs were shown in Table 3, with no significant differences (P=0.414). The overall distribution of good, medium and poor performance was 53.9% (97 instances), 46.1% (83 instances) and 0% (0 instance), respectively.

Table 3

	good(score≥4)	medium □3≤score □4□	poor□score□3□
Kimi	28□46.7%□	32□53.3%□	0□0%□
Ernie Bot	34□56.7%□	26□43.3%□	0□0%□
ChatGPT	35□58.3%□	25□41.7%□	0□0%□

2.3 Completeness

The completeness scores of the three LLMs were shown in Table 4, with no significant differences (P=0.100). The overall distribution of good, medium and poor performance was 68.9% (124 instances), 31.1% (56 instances) and 0% (0 instance), respectively.

Table 4

	good(score≥4)	medium （3≤score＜4）	poor（score＜3）
Kimi	36（60%）	24（40%）	0（0%）
Ernie Bot	47（78.3%）	13（21.7%）	0（0%）
ChatGPT	41（68.3%）	19（31.7%）	0（0%）

2.4 Clarity

The clarity scores of the three LLMs were shown in Table 5, with no significant differences (P=0.261). The overall distribution of good, medium and poor performance was 36.7% (66 instances), 60.6% (109 instances) and 2.8% (55 instances), respectively.

Table 5

	good(score≥4)	medium （3≤score＜4）	poor（score＜3）
Kimi	26（43.3%）	32（53.3%）	2（3.3%）
Ernie Bot	24（40%）	34（56.7%）	2（3.3%）
ChatGPT	16（26.7%）	43（71.7%）	1（1.7%）

2.5 Reliability

The reliability scores of the three LLMs were shown in Table 6, with significant difference (P＜0.001). The reliability of ERNIE Bot is better, with a good performance rate of 55%. The overall distribution of good, medium and poor performance was 36.1% (65 instances), 56.7% (102 instances) and 7.2% (13 instances), respectively.

Table 6

	good(score≥4)	medium （3≤score＜4）	poor（score＜3）
Kimi	18（30.0%）	41（68.3%）	1（1.7%）
Ernie Bot	33（55.0%）	24（40%）	3（5%）
ChatGPT	14（23.3%）	37（61.7%）	9（15%）

3. The performance of three LLMs in four different domains (definition and symptoms, diagnosis, treatment, prevention).

3.1 The performance on responses of the definition and symptoms part was shown in Table 7, with no statistical difference among the three LLMs (p=0.140).

Table 7

	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%）

3.2 The performance on responses of the diagnosis part was shown in the Table 8. There is a statistical difference(p=0.0.028), and ERNIE Bot performs the best.

Table 8

	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%）

3.3 The performance on responses of the treatment part was shown in Table 9, with no statistical difference among the three LLMs (p=0.349).

Table 9

	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%）

3.4 The performance on responses of the prevention part was shown in Table 10, with no statistical difference among the three LLMs (p=0.346).

Table 10

	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%）

4）Output stability evaluation

Based on the assessment of one of the physicians, the stability of the three responses from each LLM for every question was assessed. The results showed that the outputs from Kimi and Ernie Bot were stable, while the outputs from ChatGPT were different (p=0.014), indicating a lack of stability.

Discussion

Overall, the responses of LLMs to H. pylori related medical counseling in Chinese were acceptable, providing valuable medical guidance to general public. However, their performance was not entirely satisfactory, which may be linked to language factors. At present, most LLMs are still designed based on English, receiving more training in English than in other languages. Studies demonstrated that LLMs tend to perform better in English comprehension and output compared to Chinese in medical practice[6, 10]. Therefore, we speculate that LLMs designed based on Chinese language may have more advantages in Chinese medical counseling. This study showed that LLMs developed by Chinese companies outperformed GhatGPT in several aspects.

LLMs can potentially reduce costs while maintaining patient satisfaction and medical efficiency, and they have broad application prospects in 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 relevant content and provide a more comprehensive answer. However, excessive information can lead to a decline in logical coherence and clarity. As showed in this study, LLMs performed unsatisfactorily in terms of clarity, indicating that AI's linguistic competence is still not on pair 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 reliability. Due to the complexity of AI language models, interpretability and transparency are difficult to ensure. Online health information sometimes 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 where LLM outputs were “imagined” without literature and data support, known as AI hallucinations. How to identify and restrict hallucinations remains a significant concern for professionals. Reasonable policies and regulations are necessary to ensure the 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.

In conclusion, LLMs demonstrate great potential in medical counseling, although professional review and further evaluation are still required.

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. The Lancet. Public Health. 2021 Dec;6(12):e888-e896. DOI: 10.1016/s2468-2667(21)00164-x. PMID: 34838195; PMCID: PMC8646857 .

[2] FitzGerald R, Smith SM. An Overview of Helicobacter pylori Infection. Methods Mol Biol. 2021;2283:1-14. doi: 10.1007/978-1-0716-1302-3_1. PMID: 33765303 .

[3] Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, Bray F. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA Cancer J Clin. 2021 May;71(3):209-249. doi: 10.3322/caac.21660. Epub 2021 Feb 4. PMID: 33538338 .

[4] Omiye JA, Gui H, Rezaei SJ, Zou J, Daneshjou R. Large Language Models in Medicine: The Potentials and Pitfalls : A Narrative Review. Ann Intern Med. 2024 Feb;177(2):210-220. doi: 10.7326/M23-2772. Epub 2024 Jan 30. PMID: 38285984 .

[5] Liu PR, Lu L, Zhang JY, Huo TT, Liu SX, Ye ZW. Application of Artificial Intelligence in Medicine: An Overview. Curr Med Sci. 2021 Dec;41(6):1105-1115. doi: 10.1007/s11596-021-2474-3. Epub 2021 Dec 6. PMID: 34874486; PMCID: PMC8648557 .

[6] Kong QZ, Ju KP, Wan M, Liu J, Wu XQ, Li YY, Zuo XL, Li YQ. Comparative analysis of large language models in medical counseling: A focus on Helicobacter pylori infection. Helicobacter. 2024 Jan-Feb;29(1):e13055. doi: 10.1111/hel.13055. PMID: 39078641 .

[7] Hatem R, Simmons B, Thornton JE. A Call to Address AI "Hallucinations" and How Healthcare Professionals Can Mitigate Their Risks. Cureus. 2023 Sep 5;15(9):e44720. doi: 10.7759/cureus.44720. PMID: 37809168; PMCID: PMC10552880 .

[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. DOI:10.3760/cma.j.cn311367-20220206-00057 .

[9] Malfertheiner P, Megraud F, Rokkas T, Gisbert JP, Liou JM, Schulz C, Gasbarrini A, Hunt RH, Leja M, O'Morain C, Rugge M, Suerbaum S, Tilg H, Sugano K, El-Omar EM; European Helicobacter and Microbiota Study group. Management of Helicobacter pylori infection: the Maastricht VI/Florence consensus report. Gut. 2022 Aug 8:gutjnl-2022-327745. doi: 10.1136/gutjnl-2022-327745. Epub ahead of print. PMID: 35944925 .

[10] 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 Sep;177:105173. doi: 10.1016/j.ijmedinf.2023.105173. Epub 2023 Aug 4. PMID: 37549499 .

[11] Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nat Med. 2023 Aug;29(8):1930-1940. doi: 10.1038/s41591-023-02448-8. Epub 2023 Jul 17. PMID: 37460753 .

[12] Zhui L, Fenghe L, Xuehu W, Qining F, Wei R. Ethical Considerations and Fundamental Principles of Large Language Models in Medical Education: Viewpoint. J Med Internet Res. 2024 Aug 1;26:e60083. doi: 10.2196/60083. PMID: 38971715; PMCID: PMC11327620 .

[13] Mahowald K, Ivanova AA, Blank IA, Kanwisher N, Tenenbaum JB, Fedorenko E. Dissociating language and thought in large language models. Trends Cogn Sci. 2024 Jun;28(6):517-540. doi: 10.1016/j.tics.2024.01.011. Epub 2024 Mar 19. PMID: 38508911 .

[14] Daraz L, Morrow AS, Ponce OJ, Farah W, Katabi A, Majzoub A, Seisa MO, Benkhadra R, Alsawas M, Larry P, Murad MH. Readability of Online Health Information: A Meta-Narrative Systematic Review. Am J Med Qual. 2018 Sep/Oct;33(5):487-492. doi: 10.1177/1062860617751639. Epub 2018 Jan 18. PMID: 29345143 .

[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 Mar 6;9:e46885. doi: 10.2196/46885. PMID: 36863937; PMCID: PMC10028514 .

[16] Aljamaan F, Temsah MH, Altamimi I, Al-Eyadhy A, Jamal A, Alhasan K, Mesallam TA, Farahat M, Malki KH. Reference Hallucination Score for Medical Artificial Intelligence Chatbots: Development and Usability Study. JMIR Med Inform. 2024 Jul 31;12:e54345. doi: 10.2196/54345. PMID: 39083799; PMCID: PMC11325115 .

[17] Athaluri SA, Manthena SV, Kesapragada VSRKM, Yarlagadda V, Dave T, Duddumpudi RTS. Exploring the Boundaries of Reality: Investigating the Phenomenon of Artificial Intelligence Hallucination in Scientific Writing Through ChatGPT References. Cureus. 2023 Apr 11;15(4):e37432. doi: 10.7759/cureus.37432. PMID: 37182055; PMCID: PMC10173677 .