

Assessing the performance of large language models (LLMs) in answering medical questions regarding breast cancer in the Chinese context

DIGITAL HEALTH
Volume 10: 1–11
© The Author(s) 2024
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/20552076241284771
journals.sagepub.com/home/dhj



Ying Piao^{1,†} , Hongtao Chen^{1,†}, Shihai Wu¹, Xianming Li¹, Zihuang Li¹
and Dong Yang¹

Abstract

Purpose: Large language models (LLMs) are deep learning models designed to comprehend and generate meaningful responses, which have gained public attention in recent years. The purpose of this study is to evaluate and compare the performance of LLMs in answering questions regarding breast cancer in the Chinese context.

Material and Methods: ChatGPT, ERNIE Bot, and ChatGLM were chosen to answer 60 questions related to breast cancer posed by two oncologists. Responses were scored as comprehensive, correct but inadequate, mixed with correct and incorrect data, completely incorrect, or unanswered. The accuracy, length, and readability among answers from different models were evaluated using statistical software.

Results: ChatGPT answered 60 questions, with 40 (66.7%) comprehensive answers and six (10.0%) correct but inadequate answers. ERNIE Bot answered 60 questions, with 34 (56.7%) comprehensive answers and seven (11.7%) correct but inadequate answers. ChatGLM generated 60 answers, with 35 (58.3%) comprehensive answers and six (10.0%) correct but inadequate answers. The differences for chosen accuracy metrics among the three LLMs did not reach statistical significance, but only ChatGPT demonstrated a sense of human compassion. The accuracy of the three models in answering questions regarding breast cancer treatment was the lowest, with an average of 44.4%. ERNIE Bot's responses were significantly shorter compared to ChatGPT and ChatGLM ($p < .001$ for both). The readability scores of the three models showed no statistical significance.

Conclusions: In the Chinese context, the capabilities of ChatGPT, ERNIE Bot, and ChatGLM are similar in answering breast cancer-related questions at present. These three LLMs may serve as adjunct informational tools for breast cancer patients in the Chinese context, offering guidance for general inquiries. However, for highly specialized issues, particularly in the realm of breast cancer treatment, LLMs cannot deliver reliable performance. It is necessary to utilize them under the supervision of healthcare professionals.

Keywords

Breast cancer, large language model, ChatGPT, ERNIE Bot, ChatGLM, Chinese context

Submission date: 17 January 2024; Acceptance date: 3 September 2024

¹Department of Radiation Oncology, Shenzhen People's Hospital (The Second Clinical Medical College, Jinan University; The First Affiliated Hospital, Southern University of Science and Technology), Shenzhen, Guangdong, People's Republic of China

[†]Ying Piao and Hongtao Chen are first authors.

Corresponding author:

Yang Dong, Department of Radiation Oncology, Shenzhen People's Hospital (The Second Clinical Medical College, Jinan University; The First Affiliated Hospital, Southern University of Science and Technology), Dongmenbei Road 1017, Shenzhen 518000, Guangdong, People's Republic of China. Email: yangdongyifan888@sina.com



Introduction

Large Language Models (LLMs) are deep learning models designed to comprehend textual information and generate meaningful responses.¹ ChatGPT (Chat Generative Pre-trained Transformer),² which was released by OpenAI on 30 November 2022, is a deep learning LLM that has received global attention from academic communities across a wide range of fields because of its exceptional ability to provide near-human-quality answers. At present, many Chinese technology companies and research institutions have also released multiple LLM products, such as ERNIE Bot developed by Baidu,³ BlueLM developed by vivo,⁴ Qwen developed by Alibaba,⁵ ChatGLM developed by Tsinghua University,⁶ MOSS developed by Fudan University,⁷ and so on. Due to their extensive training on large amounts of Chinese datasets, these LLMs might establish a competitive advantage and long-term potential for improvement in the Chinese context.

LLMs, with either direct conversation or as part of a search engine, are rapidly becoming an important source for patients to access medical-related information, with potential applications including explaining medical reports,^{8,9} providing recommendations relating to cancer prevention, screening, diagnosis, treatment, and follow-up,^{10–15} offering advice for perioperative care,¹⁶ and so on. According to a report published by Pew Research Center in 2013,¹⁷ 60% of Americans seek medical information online, primarily for self-diagnosis or to obtain more information about specific diseases. In China, people also have a similar habit of searching for medical-related information using search engines like Baidu or other online platforms.

With the development of LLM, higher-quality information can be obtained more quickly and conveniently, especially in the medical field. These suggestions provided by LLM might significantly influence user's medical decisions. Hence, research from healthcare professionals is necessary for ensuring the accuracy and quality of the answers provided by LLM in the medical domain. To date, most published research^{9–16} has focused on exploring the accuracy, completeness, and reproducibility of LLMs, such as ChatGPT and Google Bard,¹⁸ in answering medical questions in English. However, there is a lack of experience in applying it to languages other than English. Different languages possess distinct complex systems, which reflect the cultural characteristics of diverse ethnic groups. Despite Chinese being one of the most widely spoken languages in terms of the number of users, different conclusions have been reached when comparing ChatGPT's capability in the English and Chinese contexts.^{19,20}

Breast cancer has the highest incidence rate among female malignant tumors in both USA and China.^{21,22} Thanks to the implementation of early cancer screening

and the development of medical technology, the 5-year survival rate of breast cancer patients has reached 80%–90%.^{21,23} Therefore, in both English and Chinese contexts, information and discussion about breast cancer online is abundant. So, we chose breast cancer as an example to evaluate and compare the ability of LLMs in answering medical-related questions in the Chinese context.

Methods

LLM selection

Given that ChatGPT has the largest user base worldwide, this study takes ChatGPT as the baseline. ERNIE Bot was reported to be the most widely used LLM developed in Chinese.²⁴ ChatGLM, developed with the participation of China's top university, Tsinghua University, is the earliest open-source LLM in China. The GLM-130B LLM was open-sourced in August 2022. So, the two LLMs were chosen as the LLM testing samples designed explicitly for the Chinese language context. In this study, the ChatGPT model was based on version GPT-3.5-Turbo, the ERNIE Bot on version 2.2.3, and the ChatGLM on version 2.0.

Question and response generation

Sixty questions (Table 1) regarding breast cancer prevention (8, 13.3%), diagnosis (15, 25.0%), treatment (27, 45.0%), and follow-up (10, 16.7%) were created by two working senior oncologists, both having more than 10 years of experience in breast tumor care. Most of the questions are frequently asked by breast cancer patients. Then the Chinese questions were separately entered into the three LLMs as text, without being translated into English, and the dialogue was restarted after each answer generation without providing any feedback. All the queries are expressed in Mandarin Chinese, rather than regional dialect. The responses generated were originally recorded in Chinese. For better reader comprehension, only the translated English version is provided in Supplemental Table S1 (anyone who wishes to view the original Chinese version can request it via email).

Grading

Reviewing and grading of each answer were done independently by two experienced oncologists. The accuracy of each response was graded using the following grading system: 1. comprehensive, 2. correct but inadequate, 3. mixed with correct and incorrect data, 4. completely incorrect, and 5. unanswered. Discrepancies in grading between the two reviewers were independently reviewed and resolved by a third oncology expert. We graded the

Table 1. Questions posed to the three language models, with the results of the assessments of the answers.

Questions	Score		
	ChatGPT	ERNIE Bot	ChatGLM
Prevention	1. What is breast cancer?	1	1
	2. Who is more likely to get breast cancer?	1	2
	3. Can I prevent breast cancer?	1	1
	4. I'm afraid I might have breast cancer, what tests should I undergo?	1	3
	5. What are the symptoms of breast cancer?	1	1
	6. Do I have a higher risk of developing breast cancer if my aunt has diagnosed with breast cancer?	1	1
	7. How often should I undergo breast ultrasound screenings if I don't have any symptoms of breast cancer?	1	1
	8. Why do I develop breast cancer?	1	1
Diagnosis	9. Which one is better, breast ultrasound scan or mammograms ?	1	1
	10. Will cystic breast lesions detected on ultrasound transform into breast cancer?	1	1
	11. Will breast fibroadenoma detected on ultrasound transform into breast cancer?	1	1
	12. Will breast hyperplasia detected on ultrasound transform into breast cancer?	1	1
	13. My report says calcifications on breast mammography, what does that mean?	1	1
	14. Should all breast cancer patients undergo MRI scans?	1	3
	15. My breast ultrasound report says I have a BI-RADS 2a lesion. What does this mean?	4	4
	16. My breast ultrasound report says I have a BI-RADS 4d lesion. What does this mean?	4	4
	17. My breast ultrasound report says I have a BI-RADS 7 lesion. What does this mean?	4	4
	18. My breast mammogram report says I have a BI-RADS 3 lesion. What does this mean?	1	3
	19. Why do I need to undergo a breast biopsy for pathological examination ?	1	1
	20. Can breast biopsy lead to tumor spreading?	1	1
	21. My pathology report says I have high-grade DCIS in breast. What does this mean?	1	3
	22. My pathology report says I have invasive ductal carcinoma, Nottingham grade 2. What does this mean?	3	1
	23. My pathology report says I have breast cancer with lymph node micrometastasis. What does this mean?	2	3
Treatment	24. I have been diagnosed with breast cancer. Should I choose breast-conserving surgery or radical surgery?	3	1
	25. Does genetic testing for breast cancer provide helpful information for diagnosis and treatment?	1	1
	26. What are the molecular subtypes of breast cancer?	3	3
	27. What are the treatment options for early-stage breast cancer?	1	1
	28. What are the treatment options for locally advanced breast cancer?	1	1
	29. Is stage IV breast cancer with lung metastasis curable?	1	3

(continued)

Table 1. Continued.

Questions	Score		
	ChatGPT	ERNIE Bot	ChatGLM
30. Is postoperative treatment necessary after breast cancer surgery?	1	3	1
31. What are the potential symptoms or long-term effects after breast cancer surgery?	1	1	3
32. What are the common chemotherapy regimens for breast cancer?	3	2	2
33. What are common side effects of breast cancer chemotherapy?	1	2	1
34. Is radiation therapy necessary after breast cancer surgery?	1	3	2
35. What is the recommended radiation therapy dose for breast cancer?	3	3	3
36. What are common side effects of breast cancer radiotherapy?	1	1	2
37. What are targeted therapy drugs for breast cancer ?	3	3	3
38. Is TDM1 treatment necessary for breast cancer patients?	3	3	3
39. What are common side effects of breast cancer target therapy?	1	1	3
40. Is metronomic chemotherapy necessary for breast cancer patients?	3	3	5
41. Can breast cancer patients consider breast reconstruction surgery as a treatment option?	3	1	1
42. Is endocrine therapy necessary after breast cancer surgery?	1	1	1
43. Is ovarian function suppression therapy necessary after breast cancer surgery?	2	3	2
44. Is Immunotherapy necessary after breast cancer surgery?	2	2	3
45. Can breast cancer be treated solely with traditional Chinese medicine?	1	1	1
46. Is stage IV breast cancer with brain metastasis curable?	1	1	1
47. Is stage IV breast cancer with bone metastasis curable?	2	2	2
48. A 69-year-old female patient diagnosed with invasive ductal carcinoma of the breast, T2N0M0, ER+, PR+, 3 Her2+, Fish-, Ki-67 10%+, what treatment approach should be taken for?	3	2	3
49. A 61-year-old female patient diagnosed with ductal carcinoma in situ (DCIS) of the breast, TisN0M0, ER+, PR +, Her2-, Ki-67 5%+, what treatment approach should be taken for?	1	1	3
50. A 50-year-old female patient diagnosed with invasive ductal carcinoma of the breast, T4N3M0, ER-, PR-, 2 Her2+++, Ki-67 50%+, what treatment approach should be taken for?	2	3	3
Follow-up 51. Can breast cancer patients continue working normally?	1	1	1
52. Can you introduce the follow-up plan after breast cancer treatment?	2	3	1
53. Is it possible for breast cancer patients to conceive or have children?	1	1	1
54. Can breast cancer patients breastfeed?	3	1	1
55. Can breast cancer patients engage in normal sexual activity?	1	1	1
56. Is calcium supplementation necessary for breast cancer patients?	1	1	1

(continued)

Table 1. Continued.

Questions	Score		
	ChatGPT	ERNIE Bot	ChatGLM
57. Is zoledronic acid necessary for breast cancer patients?	1	1	1
58. Will endocrine therapy cause lipid abnormalities for breast cancer patients?	1	1	3
59. Will endocrine therapy cause osteoporosis for breast cancer patients?	1	1	1
60. Are there any dietary considerations during endocrine therapy for breast cancer?	1	2	1

Note. Each response was scored using the following grading system: 1. comprehensive, 2. correct but inadequate, 3. mixed with correct and incorrect data, 4. completely incorrect, and 5. unanswered.

responses according to the Chinese Society of Clinical Oncology (CSCO) guidelines.²⁵ The grading was conducted in a blinded and independent manner, where the reviewers were unaware of the model used for each question and the scores of others, including the third reviewer.

Readability assessment

Readability for each response was assessed using natural language recognition programming. The words in responses were classified into different difficulty levels in ascending order (1, 2, 3, and 4) based on the *General Standard Chinese Character Table* and the *Syllabus of Graded Words and Characters for Chinese Proficiency*, readability databases used in Chinese linguistic research. Consistent with previous studies,^{26,27} readability scores were calculated using the formula: $Y = 24.345 \times (\text{percentage of Word Level 2}) + 10.153 \times (\text{percentage of Word Level 3}) + 25.150 \times (\text{percentage of Word Level 4}) + 0.091 \times (\text{average character numbers in sentences})$. A lower score indicates better readability.

Statistical analysis

The number of responses with scores: 1. comprehensive, 2. correct but inadequate, 3. mixed with correct and incorrect data, 4. completely incorrect, and 5. unanswered was determined.

1. The proportion of different scores in the responses was calculated separately for the four groups (prevention, diagnosis, treatment, and follow-up) of answers from the three LLMs.
2. Paired *t*-test was used to compare the scores between each pair of two groups of answers.
3. Chi-square test was performed to test the accuracy among ChatGPT, ERNIE Bot, and ChatGLM, as

follows: (a) comprehensive answers versus others and (b) comprehensive or correct but inadequate answers versus others. The dichotomized scores were subjected to chi-square test.

4. Additionally, the word count for each response was calculated, and a paired *t*-test was conducted.
5. The readability scores of the three models were also compared using a paired *t*-test.

All analyses were performed using SPSS Statistics 22.

Results

Table 1 includes the 60 questions posed to the three LLMs, along with the results of assessments of the answers. Supplemental Table S2 presents the grading scores of each reviewer. ChatGPT answered 60 questions. Of the 60 answers, 40 (66.7%) were comprehensive, six (10.0%) were correct but inadequate, 11 (18.3%) were mixed with correct and incorrect data, and three (5.0%) were completely incorrect. ERNIE Bot answered 60 questions, with 34 (56.7%) being comprehensive, seven (11.7%) being correct but inadequate, 16 (26.7%) being mixed with correct and incorrect data, and three (5.0%) being completely incorrect. ChatGLM provided 60 answers to the questions, but one of the answers was not related with the question. The answers were rated 35 (58.3%) comprehensive, six (10.0%) correct but inadequate, 15 (25.0%) mixed with correct and incorrect data, three (5.0%) completely incorrect, and one (1.7%) unanswered (Figure. 1). Of the three LLMs, only ChatGPT demonstrated a sense of human compassion.

1. The count of comprehensive answers in response to questions about breast cancer prevention provided by ChatGPT, ERNIE Bot, and ChatGLM was eight out of eight (100%), six out of eight (75%), and seven out of eight (87.5%). For questions about diagnosis, the

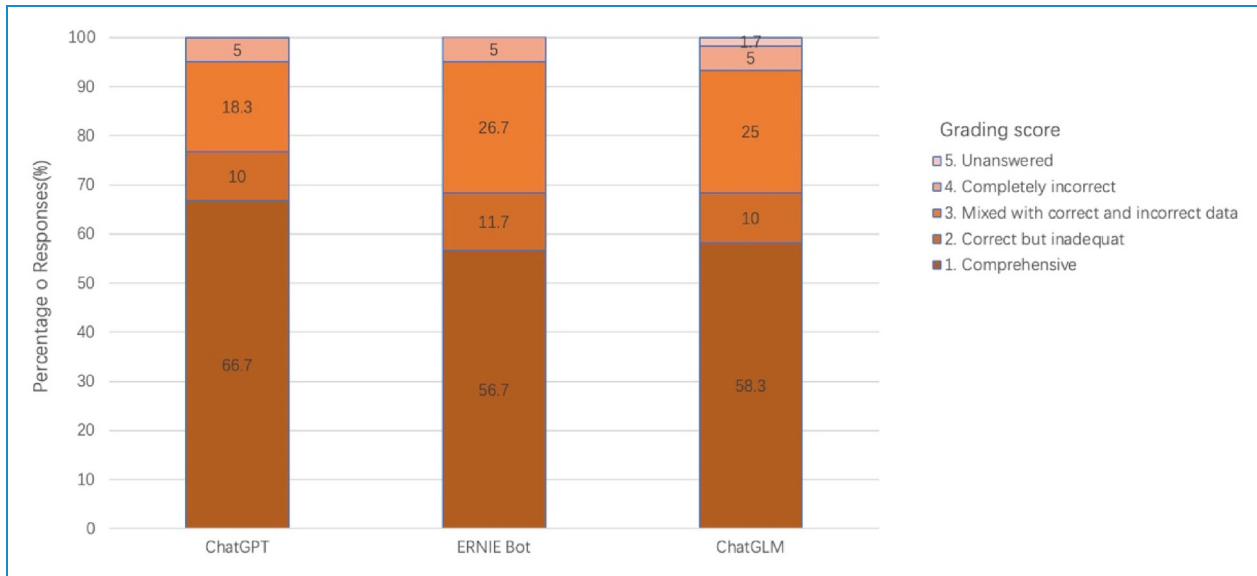


Figure 1. Percentages of response distribution using different large language models.

three LLMs provided comprehensive answers in 10 out of 15 (66.7%), eight out of 15 (53.3%), and nine out of 15 (60%) of the cases. Regarding questions about breast cancer treatment, the three LLMs provided comprehensive answers in 14 out of 27 (51.9%), 12 out of 27 (44.4%), and 10 out of 27 (37.0%) of the cases. In response to questions about follow-up, the three LLMs provided comprehensive answers in eight out of 10 (80%), eight out of 10 (80%), and nine out of 10 (90%) of the cases (Figure. 2).

2. Paired *t*-test was used to compare the scores between each pair of two groups of answers. There was no statistically significant difference between the comparisons of ChatGPT and ChatGLM ($P = .142$), ChatGPT and ERNIE Bot ($P = .132$), as well as ERNIE Bot and ChatGLM ($P = .908$).
3. Chi-square test was performed to test the accuracy among ChatGPT, ERNIE Bot, and ChatGLM, as follows:
 - (a) There were no statistically significant differences observed in the comparisons between comprehensive answers versus other answers for ChatGPT and ChatGLM ($p = .171$), ChatGPT and ERNIE Bot ($p = .118$), as well as ChatGLM and ERNIE Bot ($p = .794$).
 - (b) The comparisons between comprehensive or correct but inadequate answers versus other answers for ChatGPT and ChatGLM ($p = .165$), ChatGPT and ERNIE Bot ($p = .165$), and ChatGLM and ERNIE Bot ($P = 1.000$) did not show any statistically significant differences.
4. According to the paired *t*-test, the answers provided by ERNIE Bot (with an average of 304.2 words per response) were significantly shorter than those of ChatGPT (with an

average of 453.6 words per response) ($p < .001$) and ChatGLM (with an average of 425.0 words per response) ($p < .001$). In addition, the responses generated by ChatGPT were significantly longer than those produced by ERNIE Bot ($p < .001$), indicating a statistically significant difference. In contrast, there was no statistically significant difference in length compared to ChatGLM ($p = .11$) (Supplemental Table 3).

5. According to the paired *t*-test, the readability scores for ERNIE Bot, averaging 8.68 per response, did not show significant differences when compared with ChatGPT, which had an average score of 8.65 per response ($p = .87$), or with ChatGLM, which averaged 8.35 per response ($p = .052$). Additionally, the responses generated by ChatGPT did not differ significantly from those of ERNIE Bot ($p = .87$) or ChatGLM ($p = .08$) (Supplemental Table S4).

Discussion

Principal findings

In this study, breast cancer was chosen as an example to evaluate and compare the current capability of LLMs in answering breast cancer-related questions in the Chinese context. According to the assessment results of the three senior oncologists, ChatGPT, ERNIE Bot, and ChatGLM showed the ability to generate answers to most of the questions related to breast cancer.

According to the results, the answers had an average length of 304–454 characters, ensuring detailed responses to patient queries and including explanations of complex medical terminology. However, LLMs often mixed incorrect information

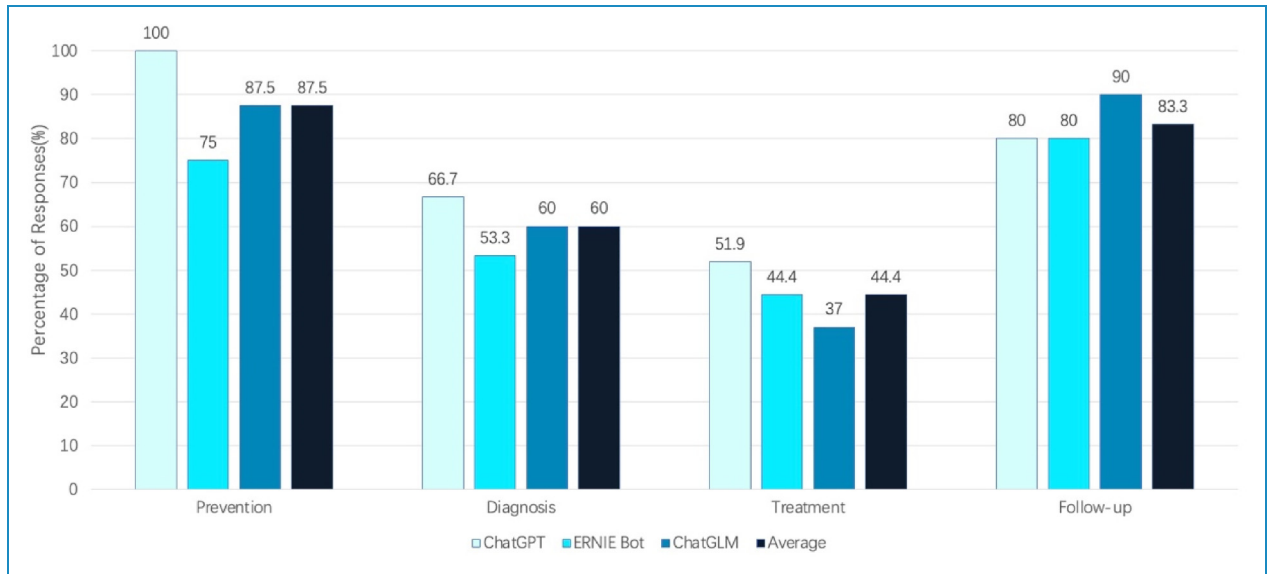


Figure 2. The proportion of comprehensive answers provided by large language models in breast cancer prevention, diagnosis, treatment, and follow-up.

with correct viewpoints when answering questions, leading to user confusion. For instance, in question 4, ERNIE Bot suggested in their response that patients who had not been diagnosed with breast cancer should undergo a full-body examination. In question 13, ChatGLM advised users with calcifications in their breast mammography to take a mammography examination for further evaluation. In question 24, ChatGPT asserted that radical surgery can ensure the complete removal of all cancer cells in the patient's body. Furthermore, the concept of endocrine therapy, targeted therapy, and immunotherapy was ambiguous. For example, in question 28, ChatGLM referred to the use of traditional Chinese medicine, such as ginsenoside rh2, to modulate the immune system as "immunotherapy." In question 37, both endocrine therapy and immune checkpoint inhibitors were called as "targeted therapy" by ERNIE Bot. In the same question, ChatGPT classified endocrine therapy drugs like tamoxifen as "targeted therapy." This phenomenon can also commonly be observed in the studies conducted in the English context.²⁸ These answers with mixed incorrect information indicate that the capability of LLMs in answering breast cancer-related questions in the Chinese context has not yet been reliable at the current stage. It is important for the LLMs to be applied under professional medical review, especially in questions requiring high expertise.

In the analysis of the results, ChatGPT showed a higher likelihood of providing comprehensive answers (66.7%), followed by ChatGLM (58.3%) and ERNIE Bot (56.7%), with no statistically significant differences. Furthermore, ChatGPT achieved the highest percentage of comprehensive or correct but inadequate answers (scores 1 or 2) (76.7%), followed by both ERNIE Bot and ChatGLM

(68.3%), without any statistical significance. Pairwise *t*-tests were also conducted to compare the accuracy and readability scores of the responses generated by the three LLMs, revealing no statistically significant differences. Based on the above results, there were no statistically significant differences in the accuracy and readability of the answers among the three LLMs, despite the differences in the length of the answers (ChatGPT > ChatGLM > ERNIE Bot). These findings indicate that, in the Chinese context, the capabilities of the three LLMs are similar in answering breast cancer-related questions at present. The comparable qualification rates demonstrate that LLMs developed in the Chinese context are as effective as those developed mainly in English in supporting breast cancer patients online. This ensures that individuals who prefer or are more comfortable with Chinese can equally benefit from the artificial intelligence (AI)-generated responses. What's more, it seems that ChatGPT exhibits insufficient multilingual adaptation and translation capabilities when responding to Chinese questions. For example, in question 32, "capecitabine" was incorrectly translated as "依托泊苷" (etoposide) by ChatGPT. This may suggest that within the current Chinese context, LLMs developed in Chinese have their own linguistic advantage. Of course, since most information online is expressed in Mandarin Chinese, when questions are posed in Cantonese or other less commonly used dialects, greater challenges might be posed for all three LLMs.²⁹

The differences in performance among the three LLMs may be attributed to several factors. One of the most critical components is the difference in model architecture. While most current LLMs typically employ the transformer

architecture, subtle differences in architectural design may have varying numbers of parameters and architectural designs. These differences can affect the model's ability to learn complex patterns and process various types of data efficiently. Other factors such as training methods, optimization algorithms, and decoding strategies also play a significant role in influencing the performance of LLMs. These elements influence how effectively the models learn from data, optimize their parameters, and generate coherent and contextually appropriate responses.^{1,30}

In addition, the quality of training data is a reason that cannot be ignored and significantly affects language model performance. Specialized information, such as the latest version of guidelines, is not as readily available as on open-source websites, resulting in less accurate and complete responses generated by language models than clinical guidelines. For example, in question 58, ChatGLM asserted that aminoglutethimide is commonly used in the endocrine treatment of breast cancer. However, it is pertinent to note that aminoglutethimide is a first-generation aromatase inhibitor. In current clinical practice, third-generation aromatase inhibitors, such as letrozole and anastrozole, are more commonly used due to their superior efficacy and safety profile. In response to question 39, ChatGLM indicated that the monthly cost of the Trastuzumab treatment was RMB 20,000, whereas the actual prevailing price is approximately RMB 6000. This discrepancy could exert considerable psychological strain on patients and potentially affect their decisions regarding treatment, given that cost considerations sometimes play a critical role in the assessment of therapeutic alternatives. The result provided above aligns with our research: when the questions were divided into four groups, it was found that the three models had the lowest level of accuracy (44.4%) in answering the treatment-related questions, which may be attributed to the specialized nature of oncology treatment and the frequent updating of the treatment section of the guidelines. A potential future solution could involve training LLMs with specific datasets, such as treatment guidelines, to handle specialized medical tasks. An illustrative example of this perspective is represented by GPT application programming interface (API), which allows developers to integrate and leverage the language generation capabilities of GPT models programmatically to enhance their applications, services, or systems. Wu et al.³¹ investigated the application of the GPT API for efficient radiation toxicity monitoring in prostate cancer patients, demonstrating its effectiveness in reducing the time demands on radiation oncologists.

When answering questions regarding breast cancer diagnosis, the three LLMs exhibited an average accuracy of only 60% (9/15), slightly higher than the average rating in treatment aspects. This is primarily because, in questions 15–17, all three LLMs provided answers for non-existent breast imaging categories, including “BI-RADS 2a,

BI-RADS 4d, and BI-RADS 7.” This phenomenon is known as “hallucination” in the realm of machine learning, which refers to the occurrence of unreal and detached results that neural network models infer from existing information.^{32,33} This is a phenomenon that is also common in the English environment.^{10,34}

As for the general questions, such as breast cancer prevention and follow-up, all three LLMs achieved an average accuracy rate of over 80%. This means that in the Chinese context, these LLMs may serve as a helpful adjunct tool for patients in addition to search engines when providing guidance on general questions. The study conducted by Haver et al., which utilized GPT as the subject of investigation, presented similar viewpoints in the English context.¹¹

Studies indicate that breast cancer is the most common malignancy among women in both China and the United States.^{21,22} In China, breast cancer tends to occur at a younger age, with an average diagnosis age approximately 10 years earlier compared to the United States and European Union countries.³⁵ Furthermore, due to the economic disparities in China, the incidence rate among urban women is significantly higher when compared to rural areas.³⁶ Therefore, breast cancer patients in China are more likely to seek breast cancer-related information through online sources. However, it is worth noting that breast cancer patients have a higher proportion of anxiety and depression, reaching up to 30%–40%,^{37,38} which has been demonstrated to potentially accelerate tumor metastasis and lead to adverse prognosis.³⁹ These negative emotions could be further exacerbated by the abundance of irrelevant, partially true, and even deceptive information and advertisements on the Internet, potentially leading to catastrophic consequences. Thus, the accuracy of medical advice is crucial for the network users.

According to our research, of the three LLMs evaluated, ChatGPT demonstrated a sense of human compassion. For instance, in questions 24 and 47, ChatGPT responded to the questioner with well-wishes, such as “I wish you a speedy recovery.” In questions 46 and 47, ChatGPT exhibited compassion to patients. This result is consistent with previous studies conducted in the English context.^{40,41} Research⁴⁰ indicated that ChatGPT outperformed physicians in conveying empathy and concern during patient interactions, qualities that are scarce in Chinese LLMs. The patient's anxiety can be alleviated by the comforting tone reflected in the responses. However, this does not imply that the role of oncologists can be replaced by LLMs. The potential ethical implications and challenges associated with using LLMs, particularly in the context of delivering medical information that intersects with privacy concerns, must not be overlooked.

LLM displays several advantages in offering medical information. First of all, LLM is trained on large-scale datasets that are typically selected and validated to ensure the

quality and accuracy, thereby minimizing false information and advertisements. Secondly, LLM provides answers in a conversational manner, making them more human-like and easily comprehensible. As a result, LLMs have exhibited their potential applications in health care, including supporting patients and medical practitioners in making informed health decisions, improving scientific writing, enhancing research equity, versatility, and efficiency, among other benefits.⁴² However, as reviewed elsewhere, LLMs also present with a set of unique challenges and limitations that are important for us to be aware of.^{42,43} For example, the risk of misinformation and distortion of scientific facts is highly dangerous and requires attention. Additionally, the data used to train ChatGPT and other major LLMs are not publicly available, making accurate validation of the information used to produce these outputs presently impossible. Moreover, legal and ethical questions concerning privacy, data security, and liability are also issues that cannot be overlooked. In our opinion, collaborations among AI experts, healthcare professionals, and regulatory bodies are needed to shape the future direction of LLM integration in healthcare. With the increasing use of LLMs in the Chinese language, they will become even more effective and sympathetic as they continue to be trained on larger and more diverse Chinese language datasets in the near future. Then, LLMs can offer significant support in improving patient satisfaction, alleviating anxiety, increasing compliance, and enhancing the quality of clinical service.

Limitations

There are several limitations to this study. GPT4, a more powerful and accurate version, has been launched by OpenAI on 3 March 2023. However, due to GPT4 being a paid version and having fewer users, our study still chose ChatGPT (GPT-3.5-Turbo), which has the highest user base, as the LLM control. Moreover, each question was input only once into the three LLMs, and as such, the reproducibility of responses was not assessed. While existing literature indicates a reproducibility rate exceeding 90% in ChatGPT's responses to tumor-related inquiries,^{11,14,15} the reproducibility of Chinese LLMs in answering medical questions has not been explored. To some extent, the accuracy and precision of the study might be affected. Future research could explore this aspect in greater depth to investigate the reproducibility of Chinese LLMs.

There are also some unavoidable limitations in conducting this study. Firstly, while the questions were created by the oncologists who tried their best to ask from patients' perspective, some sophisticated queries appeared to have been written by individuals with high level of medical background and specialized knowledge. From another perspective, it is these difficult professional questions that

challenge the models to demonstrate their expertise. Additionally, while the physicians conducted blind evaluations, the reviewers were aware that these responses were generated by LLMs. Therefore, the grading could be more stringent, and the performance of LLMs might be underestimated. Furthermore, since this study was conducted by clinicians, the questions regarding breast cancer treatment are inherently more complex, which probably result in lower scores in this part. Future studies could investigate whether the evaluation of LLM-generated text by other subspecialists in breast cancer might yield different results.

Conclusions

In summary, our study is the first to evaluate and compare the performance of different LLMs in the Chinese context, specifically utilizing breast cancer queries as the study model. In the Chinese context, the capabilities of ChatGPT, ERNIE Bot, and ChatGLM are similar in answering breast cancer-related questions at present. Of the three LLMs evaluated, ChatGPT demonstrated a sense of human compassion, which is currently lacking in Chinese LLMs. The three LLMs can be used as auxiliary information tools for breast cancer patients, providing guidance for general questions. However, for highly specialized issues, particularly in the realm of breast cancer treatment, LLMs cannot deliver reliable performance. It is necessary to utilize them under the supervision of healthcare professionals. Currently, the appropriate usage of LLM encounters numerous challenges, including limitations in domain-specific knowledge, constraints in model performance, potential ethical risks, and the risk of misuse. Given these challenges, it is imperative to implement targeted strategies, including engaging domain experts, continuously refining the model, enforcing specialized oversight, and bolstering legal regulations. Collectively, these measures pave the way for the appropriate usage of LLM, fostering its beneficial impact while mitigating associated risks.

Contributorship: YP and HC contributed equally to this work. DY, HC, and YP were responsible for the study concepts and design. HC and YP carried out the experimental studies and data analysis. SW, XL, and DY graded the answers of the three LLMs. HC carried out the statistical analysis. ZL and DY revised the manuscript. All authors read and approved the final manuscript.



Data availability: All data supporting the findings of this study are available within the paper and its supplementary information.

Declaration of conflicting interests: The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding: This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Guarantor: DY

Informed consent and ethical approval: This study is not a clinical trial nor a human trial; therefore, an IRB approval is not required. Therefore, the consent statement is not necessary.

ORCID iDs: Ying Piao  <https://orcid.org/0000-0002-6829-3538>
Dong Yang  <https://orcid.org/0009-0007-8746-7445>

Supplemental material: Supplemental material for this article is available online.

References

- Brown TB, Mann B, Ryder N, et al. Language models are few-shot learners. *Adv Neural Inf Process Syst* 2020; 33: 1877–1901.
- ChatGPT. OpenAI. Available at: <https://openai.com/blog/chatgpt>
- ERNIE Bot. Baidu. Available at: <https://yiyan.baidu.com/>
- BlueLM. Vivo. Available at: <https://developers.vivo.com/product/ai/bluelm>
- Qwen. Alibaba. Available at: <https://qianwen.aliyun.com/>
- ChatGLM. Zhipu AI. Available at: <https://chatglm.cn/detail>
- MOSS. Fudan University. Available at: <https://moss.fastnlp.top>
- Elkassam AA and Smith AD. Potential use cases for ChatGPT in radiology reporting. *Am J Roentgenol* 2023; 221: 373–376.
- Haver HL, Lin CT, Sirajuddin A, et al. Use of ChatGPT, GPT-4, and bard to improve readability of ChatGPT's answers to questions on lung cancer and lung cancer screening. *Am J Roentgenol* 2023; 221: 701–704.
- Rahsepar AA, Tavakoli N, Kim GHJ, et al. How AI responds to common lung cancer questions: ChatGPT vs Google Bard. *Radiology* 2023; 307: e230922.
- Haver HL, Ambinder EB, Bahl M, et al. Appropriateness of breast cancer prevention and screening recommendations provided by ChatGPT. *Radiology*. 2023; 307:e230424.
- Coskun B, Ocakoglu G, Yetemen M, et al. Can ChatGPT. An artificial intelligence language model, provide accurate and high-quality patient information on prostate cancer? *Urology* 2023; 180: 35–58.
- Pan A, Musheyev D, Bockelman D, et al. Assessment of artificial intelligence chatbot responses to top searched queries about cancer. *Jama Oncol* 2023; 9:1437–1440.
- Emile SH, Horesh N, Freund M, et al. How appropriate are answers of online chat-based artificial intelligence (ChatGPT) to common questions on colon cancer? *Surgery* 2023; 174: 1273–1275.
- Yeo YH, Samaan JS, Ng WH, et al. Assessing the performance of ChatGPT in answering questions regarding cirrhosis and hepatocellular carcinoma. *Clin Mol Hepatol*. 2023; 29:721–732.
- Moazzam Z, Cloyd J, Lima HA, et al. Quality of ChatGPT responses to questions related to pancreatic cancer and its surgical care. *Ann Surg Oncol* 2023; 30: 6284–6286.
- Pew Research Center. (2013). Health online 2013. Available at: <https://www.pewresearch.org/internet/2013/01/15/health-online-2013/>
- Google Bard. Available at: <https://bard.google.com>
- Wang H, Wu WZ, 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 of Med Inform* 2023; 177: 105173.
- Shao CY, Li H, Liu XL, et al. Appropriateness and comprehensiveness of using ChatGPT for perioperative patient education in thoracic surgery in different language contexts: Survey study. *Interact J Med Res* 2023; 12: e46900.
- Giaquinto AN, Sung H, Miller KD, et al. Breast cancer statistics, 2022. *CA-Cancer J Clin* 2023; 72: 524–541.
- Xia C, Dong X, Li H, et al. Cancer statistics in China and United States, 2022: Profiles, trends, and determinants. *Chinese Med J-Peking* 2022; 135: 584–590.
- Zeng H, Chen W, Zheng R, et al. Changing cancer survival in China during 2003–15: A pooled analysis of 17 population-based cancer registries. *Lancet Glob Health* 2018; 6: e555–e567.
- Available at: <https://www.chinadailyhk.com/hk/article/581017>
- Guidelines of Chinese Society of Clinical Oncology (CSCO). Breast Cancer 2022. Available at: <https://www.cSCO.org.cn/cn/index.aspx>
- Wang W, Lyu J, Li M, et al. Quality evaluation of HPV vaccine-related online messages in China: a cross-sectional study. *Hum Vacc Immunother* 2020; 17: 1089–1096.
- Bai XY, Zhang YW, Li J, et al. Online information on Crohn's disease in Chinese: an evaluation of its quality and readability. *J Digest Dis* 2019; 20: 596–601.
- Lukac S, Dayan D, Fink V, et al. Evaluating ChatGPT as an adjunct for the multidisciplinary tumor board decision-making in primary breast cancer cases. *Arch Gynecol Obstet* 2023; 308:1831–1844.
- Fu Z, Hsu YC, Chian CS, et al. Efficacy of ChatGPT in Cantonese sentiment analysis: Comparative study. *J Med Internet Res* 2024; 26: e51069.
- Hoffmann J, Borgeaud S, Mensch A, et al. Training compute-optimal large language models[J]. arXiv preprint arXiv:2203.15556, 2022.
- Wu DJ and Bibault JE. Pilot applications of GPT-4 in radiation oncology: Summarizing patient symptom intake and targeted chatbot applications. *Radiother Oncol* 2023; 190: 109978.
- Xiao Y and Wang WY. On hallucination and predictive uncertainty in conditional language generation. arXiv preprint online Mar 28, 2021. arXiv:2103.15025, 2021.
- Rohrbach A, Hendricks L A, Burns K, et al. Object hallucination in image captioning. arXiv preprint online Sep 6, 2018. arXiv:1809.02156, 2018.
- Walker HL, Ghani S, Kuemmerli C, et al. Reliability of medical information provided by ChatGPT: assessment against clinical guidelines and patient information quality instrument. *J Med Internet Res* 2023; 25: e47479.
- Song QK, Li J, Huang R, et al. Age of diagnosis of breast cancer in China: almost 10 years earlier than in the United States and the European Union. *Asian Pac J Cancer P*. 2014; 15:10021–10025.
- Zhang ML, Peng P, Wu CX, et al. Report of breast cancer incidence and mortality in China registry regions, 2008–2012. *Chin J Oncol* 2019; 41: 315–320.

37. Tao F, Xu M, Zou Q, et al. Prevalence and severity of anxiety and depression in Chinese patients with breast cancer: a systematic review and meta-analysis. *Front Psychiatry* 2023; 14: 1080413.
 38. Hashemi SM, Rafiemanesh H, Aghamohammadi T, et al. Prevalence of anxiety among breast cancer patients: a systematic review and meta-analysis. *Breast Cancer-Tokyo* 2019; 27: 166–178.
 39. Wang YH, Li JQ, Shi JF, et al. Depression and anxiety in relation to cancer incidence and mortality: a systematic review and meta-analysis of cohort studies. *Mol Psychiatry* 2019; 25: 1487–1499.
 40. Ayers JW, Poliak A, Dredze M, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *Jama Intern Med* 2023; 183:589–596.
 41. Amin S, Kawamoto CT and Pokhrel P. Exploring the ChatGPT platform with scenario-specific prompts for vaping cessation. *Tob Control* 2023: 1–3.
 42. Schukow C, Smith SC, Landgrebe E, et al. Application of ChatGPT in routine diagnostic pathology: promises, pitfalls, and potential future directions. *Adv Anat Pathol* 2023; 31: 15–21.
 43. Iannantuono GM, Bracken-Clarke D, Floudas, et al. Applications of large language models in cancer care: Current evidence and future perspectives. *Front Oncol.* 2023; 13: 1268915.
-