Risk Management and Healthcare Policy

Dovepress
Taylor & Francis Group

Open Access Full Text Article

ORIGINAL RESEARCH

# Evaluation of Three Large Language Models' Response Performances to Inquiries Regarding Post-Abortion Care in the Context of Chinese Language: A Comparative Analysis

Danyue Xue[1,2], Sha Liao[1,2]

[1]Department of Operating Room Nursing, West China second University Hospital, Sichuan University, Chengdu, Sichuan, People's Republic of China; [2]Key Laboratory of Birth Defects and Related Diseases of Women and Children (Sichuan University), Ministry of Education, Chengdu, Sichuan, People's Republic of China

Correspondence: Sha Liao, Email 1205843055@qq.com

Risk Management and Healthcare Policy downloaded from https://www.dovepress.com/
For personal use only.

**Background:** This study aimed to evaluate the response performances of three large language models (LLMs) (ChatGPT, Kimi, and Ernie Bot) to inquiries regarding post-abortion care (PAC) in the context of the Chinese language.

**Methods:** The data was collected in October 2024. Twenty questions concerning the necessity of contraception after induced abortion, the best time for contraception, choice of a contraceptive method, contraceptive effectiveness, and the potential impact of contraception on fertility were used in this study. Each question was asked three times in Chinese for each LLM. Three PAC consultants conducted the evaluations. A Likert scale was used to score the responses based on accuracy, relevance, completeness, clarity, and reliability.

**Results:** The number of responses received "good" (a mean score > 4), "average" (3 < mean score ≤ 4), and "poor" (a mean score ≤ 3) in overall evaluation was 159 (88.30%), 19 (10.57%), and 2 (1.10%). No statistically significant differences were identified in the overall evaluation among the three LLMs ($P$ = 0.352). The number of the responses evaluated as good for accuracy, relevance, completeness, clarity, and reliability were 87 (48.33%), 154 (85.53%), 136 (75.57%), 133 (73.87%), and 128 (71.10%), respectively. No statistically significant differences were identified in accuracy, relevance, completeness or clarity between the three LLMs. A statistically significant difference was identified in reliability ($P$ < 0.001).

**Conclusion:** The three LLMs performed well overall and showed great potential for application in PAC consultations. The accuracy of the LLMs' responses should be improved through continuous training and evaluation.

**Keywords:** artificial intelligence, abortion, induced, referral and consultation, delivery of health care, comparative study

## Background

Induced abortion refers to the termination of a pregnancy through a medical procedure.[1] Approximately 56 million induced abortions occur globally every year, with nearly one-third involving repeat abortions, contributing to increased morbidity and mortality from post-abortion complications.[2,3] In 1994, several countries made political commitments to address abortion-related morbidity and mortality by providing quality healthcare. In this context, post-abortion care (PAC) was widely promoted globally to offer contraceptive counseling services to women who have undergone abortions. In subsequent clinical practice, PAC has proven effective in reducing abortion-related mortality and morbidity and preventing future unintended pregnancies.[4,5]

Healthcare consulting plays a vital role in facilitating communication between patients and healthcare professionals. It helps patients better understand their health status and take appropriate measures to address health problems. PAC is a standardized healthcare procedure designed to provide health education about contraception to women who have had an induced abortion and their family members. This is achieved by implementing a standardized post-abortion service

Received: 1 April 2025
Accepted: 9 August 2025
Published: 18 August 2025



© 2025 Xue and Liao. This work is published and licensed by Dove Medical Press Limited. The full terms of this license are available at https://www.dovepress.com/terms.php and incorporate the Creative Commons Attribution – Non Commercial (unported, v4.0) License (http://creativecommons.org/licenses/by-nc/4.0/). By accessing the work you hereby accept the Terms. Non-commercial uses of the work are permitted without any further permission from Dove Medical Press Limited, provided the work is properly attributed. For permission for commercial use of this work, please see paragraphs 4.2 and 5 of our Terms (https://www.dovepress.com/terms.php).

procedure and care system in hospitals. Such systems encourage women to adopt effective contraceptive measures promptly after an abortion, thereby reducing the incidence of induced abortions and repeat abortions caused by unintended pregnancies and improving overall reproductive health.[6]

A large language model (LLM) is an artificial intelligence (AI) model developed through training on massive datasets. It can generate outputs that closely resemble human language. LLMs have been widely adopted in the healthcare sector, reflecting core values such as healthcare consultation. They can answer clinical questions, promote interactive learning, assist in generating human-like text, provide basic guidance, and explain complex concepts. In particular, LLMs hold significant potential for enhancing telemedicine by delivering timely health information and basic guidance to patients, especially when resources are limited or physical distance poses a barrier to care.[7–10]

In China, induced abortions account for nearly one-quarter of the global total, with 55.9% of these cases involving repeat abortions.[11] Abortions, particularly repeat abortions, have become a major public health concern, threatening the reproductive health of Chinese women. Women who have undergone abortions require PAC consultations to improve their understanding of contraception and family planning, reduce the incidence of repeat abortions, and mitigate the harm caused by repeated procedures. With technological advancements, LLMs have the potential to serve as a powerful auxiliary tool in PAC consultations. Compared to general disease consultations, LLMs may face limitations in PAC consultations, such as a lack of emotional care, which requires deep humanistic support. However, they can provide patients with rapid and convenient access to health management information, positively impacting public health.[12] Despite this potential, there is a lack of a comprehensive assessment of the accuracy, relevance, completeness, clarity, and reliability of LLMs in PAC consultations. This study aims to evaluate the performance of three LLMs in responding to inquiries about PAC in the context of the Chinese language. The goal is to provide technical support for improving the dissemination of contraception knowledge among Chinese women post-abortions, reducing the rate of repeat abortions, and promoting reproductive health.

# Methods

## Ethical Considerations

This study was conducted in accordance with the Declaration of Helsinki and the requirements of relevant regulations in China. Ethical approval for this study was waived by Medical Ethics Committee of West China Second University Hospital, Sichuan University [2025 Medical Scientific Research for Ethical Approval No. (M05)]. This study investigated the performance of three LLMs' responses to questions about post-abortion care. It is an evaluation on LLMs, not on patients. Therefore, informed consent from patients was waived.

## Data Source

Data collection for this study was conducted in October 2024. Three advanced LLMs were used: ChatGPT 4.0 Turbo (developed by OpenAI, United States), Kimi 2.1.4 (developed by Moonshot AI, China), and Ernie Bot 3.5 (developed by Baidu, China). These LLMs were selected based on their representativeness in technological advancement, Chinese language processing capabilities, market popularity, source diversity, and resource availability.

## Study Design

Three healthcare professionals, each with more than 10 years of work experience and accredited PAC consultant qualifications, participated in this study. The PAC consultants communicated with patients face-to-face during outpatient visits and via text messages to identify the questions patients frequently asked. Subsequently, 20 commonly asked questions were selected for this study (Table 1).

These 20 questions covered the following five areas: (1) the necessity of contraception after induced abortion; (2) the best time for contraception; (3) choice of a contraceptive method; (4) evaluation of contraceptive effectiveness; (5) the potential impact of different methods of contraception on fertility.

The 20 questions were compiled into a set. Each question was asked three times in Chinese to each of the three LLMs. The iterative prompts for user input in the three LLMs across the three iterations were identical. Each LLM's 20 responses

**Table 1** Twenty Commonly Asked Questions

| Area | Questions |
|---|---|
| The necessity of contraception after induced abortion | 1. Must contraception be initiated after an induced abortion?; |
| | 2. Is contraception needed after an induced abortion? |
| | 3. Can contraception be avoided after an induced abortion? |
| The best time for contraception | 4. Is contraception needed immediately after an induced abortion? |
| | 5. How long should contraception be used after an induced abortion? |
| | 6. How soon after an induced abortion can one start taking oral contraceptive pills? |
| | 7. Can an intrauterine device be inserted immediately after an induced abortion? |
| Choice of a contraceptive method | 8. What are the methods of contraception after an induced abortion? |
| | 9. What is the best method of contraception after an induced abortion? |
| | 10. Which method of contraception is better for me if I plan to have a baby? |
| | 11. Which method of contraception is better for me if I do not plan to have a baby? |
| | 12. What methods of contraception shall I use if I have uterine fibroids? |
| | 13. What methods of contraception should I use if I have dysmenorrhea? |
| | 14. What methods of contraception should I use if I have breast cysts? |
| Evaluation of contraceptive effectiveness | 15. What are the different types of intrauterine devices, and how do they differ? |
| | 16. Do long-acting contraceptive pills have a better effect, or do short-acting contraceptive pills have a better effect? |
| | 17. Which method of contraception is most effective? |
| The potential impact of different methods of contraception on fertility | 18. Do oral contraceptive pills affect fertility? |
| | 19. Will inserting an intrauterine device impact the ability to have a baby? |
| | 20. Which method of contraception has the least impact on the ability to have a baby? |

from the 1st, 2nd, and 3rd question-and-answer sessions were labeled "Iteration A", "Iteration B", and "Iteration C", respectively. The consistency of the LLMs' responses was evaluated based on the results of the three iterations. To prevent deviations caused by potential interferences, all interactions with ChatGPT were conducted using the same web version subscription account. All interactions with Kimi were conducted using the same mobile application account. All interactions with Ernie Bot were conducted using the same Baidu account on Ernie Bot's mobile application. Default settings were used for all model configurations. We cleared all relevant browser caches, cookies, and question-and-answer session data after each iterative test. Then, we waited 10 seconds to ensure that the system was fully reset and ready for the next iteration.

The evaluation was conducted in accordance with the "Guide for Family Planning Services After an Induced Abortion"[13] and the "Chinese Expert Consensus on the Clinical Application of Contraceptive Methods for Women".[14] Each LLM's three responses to each of the 20 questions were recorded and assigned to the three PAC consultants for evaluation. The three PAC consultants independently scored the responses on paper at different times. A Likert scale was used to evaluate the LLMs' responses based on the following five components:

1. *Accuracy*. This evaluated whether a LLM's response was completely correct free of factual errors. A score of 5 indicated the response was entirely correct with no errors. A score of 1 indicated a significant error or that the response was completely incorrect.

2. *Relevance*. This assessed whether the response directly addressed the question without deviating from the topic. A score of 5 indicated the response was fully relevant to the question. A score of 1 indicated the response was irrelevant or completely off-topic.
3. *Completeness*. This evaluated whether the response was comprehensive and covered all key points of the question. A score of 5 indicated the response was comprehensive and included all necessary information. A score of 1 indicated the response was incomplete, with key information missing.
4. *Clarity*. This assessed whether the response was clearly expressed and easy to understand. A score of 5 indicated the response was clear, with smooth language and easy to comprehend. A score of 1 indicated the response was confusing and difficult to understand.
5. *Reliability*. This evaluated whether the response was reliable and based on reputable information or logical reasoning. A score of 5 indicated a highly reliable response supported by reputable information or logical reasoning. A score of 1 indicated the response was unreliable, lacked support, or had poor logic.

An overall evaluation was performed based on the scores for the five components described above. The results of the overall evaluation were categorized into three types: (1) Good: a mean score > 4; (2) Average: 3 < mean score ≤ 4; (3) Poor: a mean score ≤ 3.

All responses were scored independently by the three PAC consultants. The scores from the PAC consultant with the longest service among the three were used to analyze the stability of the output content of each LLM when responding to the same prompts multiple times. In this study, the operational definition of "output stability" referred to whether a LLM's output content showed statistically significant differences in the PAC consultant's scores when it independently answered the same prompts multiple times (three times).

## Statistical Analysis

SPSS 27.0 was used for statistical analysis. Quantitative data were presented as mean ± standard deviation (X±SD). Qualitative data were presented as a constituent ratio. The Chi-square test was performed on the qualitative data. Analysis of variance was conducted to test variability in repeated-measures data. The α level of significance was set at 0.05. The partitioned Chi-square test was used for multiple comparisons. The Bonferroni correction was applied to reduce the probability of a Type I error.

# Results

## Overall Evaluation

The distribution of the mean scores classified as "good", "average", and "poor" for the overall evaluation of each LLM's 60 responses to the 20 questions is shown in Table 2. Of the 180 responses, 159 (88.30%) were evaluated as good, 19 (10.57%) as average, and 2 (1.10%) as poor. No statistically significant differences were found in the mean scores for the overall evaluation among the three LLMs ($\chi^2 = 4.421$, $P = 0.352$).

## Accuracy

The distribution of mean scores for accuracy, classified as "good", "average", and "poor", for each LLMs' 60 responses to the 20 questions is shown in Table 3. No statistically significant differences were found in the accuracy mean scores

**Table 2** Overall Evaluation Results

| | **Good (a Mean Score>4) n (%)** | **Average (3<Mean Score ≤4) n (%)** | **Poor (a Mean Score ≤ 3) n (%)** | **Total n (%)** |
|---|---|---|---|---|
| ChatGPT 4.0 Turbo | 53 (88.30%) | 7 (11.70%) | 0 (0.00%) | 60 (100%) |
| Kimi 2.1.4 | 53 (88.30%) | 7 (11.70%) | 0 (0.00%) | 60 (100%) |
| Ernie Bot 3.5 | 53 (88.30%) | 5 (8.30%) | 2 (3.30%) | 60 (100%) |

**Table 3** Mean Scores for Accuracy

| | Good (a Mean Score>4) n (%) | Average (3<Mean Score ≤4) n (%) | Poor (a Mean Score ≤ 3) n (%) | Total n (%) |
|---|---|---|---|---|
| ChatGPT 4.0 Turbo | 29 (48.33%) | 23 (38.33%) | 8 (13.33%) | 60 (100%) |
| Kimi 2.1.4 | 28 (46.70%) | 26 (43.30%) | 6 (10.00%) | 60 (100%) |
| Ernie Bot 3.5 | 30 (50.00%) | 24 (40.00%) | 6 (10.00%) | 60 (100%) |

**Table 4** Mean Scores for Relevance

| | Good (a Mean Score>4) n (%) | Average (3<Mean Score ≤ 4) n (%) | Poor (a Mean Score ≤ 3) n (%) | Total n (%) |
|---|---|---|---|---|
| ChatGPT 4.0 Turbo | 47 (78.30%) | 11 (18.30%) | 2 (3.30%) | 60 (100%) |
| Kimi 2.1.4 | 54 (90.00%) | 5 (8.30%) | 1 (1.70%) | 60 (100%) |
| Ernie Bot 3.5 | 53 (88.30%) | 7 (11.70%) | 0 (0.00%) | 60 (100%) |

among the three LLMs ($\chi^2 = 0.661$, $P = 0.956$). Of the 180 responses, 87 (48.33%) were evaluated as good, 73 (40.53%) as average, and 20 (11.1%) as poor.

## Relevance

The distribution of mean scores for relevance, classified as "good", "average", and "poor", for each LLM' 60 responses to the 20 questions is shown in Table 4. No statistically significant differences were found in the relevance scores among the three LLMs ($\chi^2 = 4.993224$, $P = 0.288$). Of the 180 responses, 154 (85.53%) were evaluated as good, 23 (12.77%) as average, and 3 (1.67%) as poor.

## Completeness

The distribution of mean scores for completeness, classified as "good", "average", and "poor", for each LLM' 60 responses to the 20 questions is shown in Table 5. No statistically significant differences were found in the completeness scores among the three LLMs ($\chi^2 = 7.2888$, $P = 0.121$). Of the 180 responses, 136 (75.57%) were evaluated as good, 40 (22.2%) as average, and 4 (2.23%) as poor.

## Clarity

The distribution of mean scores for clarity, classified as "good", "average", and "poor", for each LLM' 60 responses to the 20 questions is shown in Table 6. No statistically significant differences were found in the clarity scores among the three LLMs ($\chi^2 = 2.189$, $P = 0.701$). Of the 180 responses, 133 (73.87%) were evaluated as good, 46 (22.57%) as average, and 1 (0.57%) as poor.

**Table 5** Mean Scores for Completeness

| | Good (a Mean Score>4) n (%) | Average (3<Mean Score ≤ 4) n (%) | Poor (a Mean Score ≤ 3) n (%) | Total n (%) |
|---|---|---|---|---|
| ChatGPT 4.0 Turbo | 45 (75.00%) | 14 (23.30%) | 1 (1.70%) | 60 (100%) |
| Kimi 2.1.4 | 51 (85.00%) | 9 (15.00%) | 0 (0.00%) | 60 (100%) |
| Ernie Bot 3.5 | 40 (66.70%) | 17 (28.30%) | 3 (5.00%) | 60 (100%) |

**Table 6** Mean Scores for Clarity

| | Good (a Mean Score>4) n (%) | Average (3<Mean Score ≤ 4) n (%) | Poor (a Mean Score ≤ 3) n (%) | Total n (%) |
|---|---|---|---|---|
| ChatGPT 4.0 Turbo | 44 (73.30%) | 16 (26.70%) | 0 (0.00%) | 60 (100%) |
| Kimi 2.1.4 | 44 (73.30%) | 16 (26.70%) | 0 (0.00%) | 60 (100%) |
| Ernie Bot 3.5 | 45 (75.00%) | 14 (23.30%) | 1 (1.70%) | 60 (100%) |

**Table 7** Mean Scores for Reliability

| | Good (a Mean Score>4) n (%) | Average (3<Mean Score ≤ 4) n (%) | Poor (a Mean Score ≤ 3) n (%) | Total n (%) |
|---|---|---|---|---|
| ChatGPT 4.0 Turbo | 33 (55.00%) | 27 (45.00%) | 0 (0.00%) | 60 (100%) |
| Kimi 2.1.4 | 50 (83.30%) | 9 (15.00%) | 1 (1.70%) | 60 (100%) |
| Ernie Bot 3.5 | 45 (75.00%) | 11 (18.30%) | 4 (6.70%) | 60 (100%) |

## Reliability

The distribution of mean scores for reliability, classified as "good", "average", and "poor", for each LLM' 60 responses to the 20 questions is shown in Table 7. A statistically significant difference was found in the reliability scores among the three LLMs ($\chi^2 = 21.204$, $P < 0.001$).

Pairwise comparison results showed a statistically significant difference in reliability scores between ChatGPT and Kimi. The proportion of responses rated as "good" in the reliability evaluation for Kimi was significantly higher than for ChatGPT. No statistically significant differences were found in the reliability scores between Ernie Bot and Kimi (ChatGPT versus Kimi: $\chi^2 = 13.482$, $P < 0.001$; ChatGPT versus Ernie Bot: $\chi^2 = 12.583$, $P = 0.002$; Ernie Bot versus Kimi: $\chi^2 = 2.263$, $P = 0.323$). Of the 180 responses, 128 (71.10%) were evaluated as good, 47 (26.11%) as average, and 5 (2.80%) as poor.

## Output Stability

The output stability of the LLMs' responses was evaluated based on a PAC consultant's scores for each LLM's three responses to each of the 20 questions. No statistically significant differences were found in the response scores among ChatGPT's three responses for each question ($F = 0.898$, $P = 0.413$), indicating no significant internal variability found in ChatGPT's responses. In contrast, statistically significant differences were found in the response scores among Kimi's and Ernie Bot's three responses to each question (Kimi: $F = 25.042$, $P < 0.001$; Ernie Bot: $F = 6.784$, $P < 0.001$), indicating significant internal variability in Kimi's and Ernie Bot's responses.

## LLMs' Responses to Questions Concerning Five Areas

### Necessity of Contraception After Induced Abortion

No statistically significant differences were found in the mean scores for responses to questions about the necessity of contraception after induced abortion among the three LLMs ($\chi^2 = 1.025$, $P = 0.599$) (Table 8).

**Table 8** Mean Scores for Large Language Models' Responses to Questions About the Necessity of Contraception After Induced Abortion

| | Good (a Mean Score>4) n (%) | Average (3<Mean Score ≤ 4) n (%) | Poor (a Mean Score ≤ 3) n (%) | Total n (%) |
|---|---|---|---|---|
| ChatGPT 4.0 Turbo | 26 (96.33%) | 1 (3.67%) | 0 (0.00%) | 27 (100%) |
| Kimi 2.1.4 | 26 (96.33%) | 1 (3.67%) | 0 (0.00%) | 27 (100%) |
| Ernie Bot 3.5 | 27 (100.00%) | 0 (0.00%) | 0 (0.00%) | 27 (100%) |

**Table 9** Mean Scores for Large Language Models' Responses to Questions About the Best Time for Contraception

| | Good (a Mean Score>4) n (%) | Average (3<Mean Score ≤ 4) n (%) | Poor (a Mean score ≤ 3) n (%) | Total n (%) |
|---|---|---|---|---|
| ChatGPT 4.0 Turbo | 16 (44.44%) | 19 (52.78%) | 1 (2.78%) | 36 (100%) |
| Kimi 2.1.4 | 31 (86.11%) | 5 (13.89%) | 0 (0.00%) | 36 (100%) |
| Ernie Bot 3.5 | 23 (63.89%) | 7 (19.44%) | 6 (16.67%) | 36 (100%) |

## Best Time for Contraception

A statistically significant difference was found in the mean scores for responses to questions about the best time for contraception among the three LLMs ($\chi^2 = 24.782$, $P < 0.001$) (Table 9). Pairwise comparison results showed that Kimi's responses to questions about the best time for contraception were significantly better than those of ChatGPT. Statistically significant differences were found in the means scores for responses to questions about the best time for contraception between ChatGPT and Ernie Bot or between Ernie Bot and Kimi (ChatGPT versus Kimi: $\chi^2 = 13.954$, $P = 0.001$; ChatGPT versus Ernie Bot: $\chi^2 = 10.366$, $P = 0.006$; Ernie Bot versus Kimi: $\chi^2 = 7.515$, $P = 0.023$).

## Choice of a Contraceptive Method

No statistically significant differences were found in the mean scores for responses to questions about choice of a contraceptive method among the three LLMs ($\chi^2 = 5.648$, $P = 0.227$) (Table 10).

## Contraceptive Effectiveness

A statistically significant difference was found in the responses to questions about contraceptive effectiveness among the three LLMs ($\chi^2 = 6.389$, $P = 0.041$) (Table 11).

Pairwise comparison results showed that ChatGPT's responses to questions about contraceptive effectiveness were significantly better than those of Kimi. Statistically significant differences were found in the mean scores for responses to questions about contraceptive effectiveness between ChatGPT and Ernie Bot and between Kimi and Ernie Bot (ChatGPT versus Kimi: $\chi^2 = 6.750$, $P = 0.009$; ChatGPT versus Ernie Bot: $\chi^2 = 4.320$, $P = 0.038$; Ernie Bot versus Kimi: $\chi^2 = 0.491$, $P = 0.484$).

**Table 10** Mean Scores for Large Language Models' Responses to Questions About Choice of a Contraceptive Method

| | Good (a Mean Score>4) n (%) | Average (3<Mean Score ≤ 4) n (%) | Poor (a Mean Score ≤ 3) n (%) | Total n (%) |
|---|---|---|---|---|
| ChatGPT 4.0 Turbo | 55 (87.30%) | 8 (12.70%) | 0 (0.00%) | 63 (100%) |
| Kimi 2.1.4 | 57 (90.48%) | 4 (6.35%) | 2 (3.17%) | 63 (100%) |
| Ernie Bot 3.5 | 55 (87.30%) | 8 (12.70%) | 0 (0.00%) | 63 (100%) |

**Table 11** Mean Scores for Large Language Models' Responses to Questions About Contraceptive Effectiveness

| | Good (a Mean Score>4) n (%) | Average (3<Mean Score ≤ 4) n (%) | Poor (a Mean Score ≤ 3) n (%) | Total n (%) |
|---|---|---|---|---|
| ChatGPT 4.0 Turbo | 27 (100.00%) | 0 (0.00%) | 0 (0.00%) | 27 (100%) |
| Kimi 2.1.4 | 21 (77.78%) | 6 (22.22%) | 0 (0.00%) | 27 (100%) |
| Ernie Bot 3.5 | 23 (85.19%) | 4 (14.81%) | 0 (0.00%) | 27 (100%) |

**Table 12** Mean Scores for Large Language Models' Responses to Questions About Potential Impact of Different Methods of Contraception on Fertility

| | **Good (a Mean Score>4) n (%)** | **Average (3<Mean Score ≤ 4) n (%)** | **Poor (a Mean Score ≤ 3) n (%)** | **Total n (%)** |
|---|---|---|---|---|
| ChatGPT 4.0 Turbo | 27 (100.00%) | 0 (0.00%) | 0 (0.00%) | 27 (100%) |
| Kimi 2.1.4 | 24 (88.89%) | 3 (11.11%) | 0 (0.00%) | 27 (100%) |
| Ernie Bot 3.5 | 20 (74.07%) | 7 (25.93%) | 0 (0.00%) | 27 (100%) |

### Potential Impact of Different Methods of Contraception on Fertility

A statistically significant difference was found in the responses to questions about the potential impact of different methods of contraception on fertility among the three LLMs ($\chi^2 = 8.442$, $P = 0.015$) (Table 12)

Pairwise comparison results showed that ChatGPT's responses to questions about the potential impact of different methods of contraception on fertility were better than Ernit Bot's. However, no statistically significant differences were found between ChatGPT and Kimi or between Kimi and Ernit Bot (ChatGPT versus Kimi: $\chi^2 = 3.716$, $P = 0.075$; ChatGPT versus Ernie Bot: $\chi^2 = 8.043$, $P = 0.005$; Ernie Bot versus Kimi: $\chi^2 = 1.964$, $P = 0.161$).

## Discussion

In this study, we investigated the performance of three LLMs' responses to inquiries about PAC in the context of the Chinese language. The results showed that 88.3% of their responses were evaluated as good overall. Under the supervision of healthcare professionals, LLMs could serve as an effective supporting tool in PAC consultations.

This study found that the reliability of Kimi and Ernie Bot was significantly higher than that of ChatGPT (ChatGPT versus Kimi: $\chi^2 = 13.482$, $P<0.001$; ChatGPT versus Ernie Bot: $\chi^2 = 12.583$, $P=0.002$). This may be because the study was conducted in the context of the Chinese language, which could have affected ChatGPT's performance, thereby reducing the reliability of its responses. Su et al[15] have demonstrated that ChatGPT outperforms Ernie Bot in healthcare consulting within an English-language context. However, it can be inferred that Chinese-based LLMs may have advantages in healthcare consulting in the Chinese language. This suggests that LLMs can be customized to meet the needs and regulations of specific regions.

This study also found that the three LLMs had high relevance scores, indicating that their responses were generally relevant to the questions. However, if the responses provided by the LLMs are overly repetitive in relation to the question, the amount of information output may decrease. This could affect the completeness and clarity of the information, possibly due to the way the questions were phrased or the number of questions asked.

This study revealed shortcomings in the accuracy of LLMs. During the evaluation, we observed that some LLM output contained fictional or "imagination" content not supported by reference literature. This phenomenon, known as "artificial intelligence hallucinations", occurs when LLMs generate responses that appear plausible but are false, inconsistent, or fictional, including those based on fabricated data or incorrect citations. Instances of hallucinations were identified across various models in our study. For example, when asked how long after stopping short-acting oral contraceptive pills a woman could try for a baby, one LLM responded that the woman could try immediately after stopping the pills. Similarly when asked to compare short-acting and long-acting contraceptive pills, an LLM stated that short-acting contraceptive pills were suitable for all women needing contraception without analyzing specific populations. LLMs' performance can improve through continuous training in specific healthcare domains; however, they face challenges in clinical practice, such as the inability to cite reliable references and the risk of generating unpredictable, fictional information. Although LLMs perform well in many areas, AI hallucinations remain a serious concern in healthcare, where patients who cannot verify the information may be misled into make incorrect decisions, potentially leading to serious consequences.[16,17]

This study shows that, in terms of output stability, ChatGPT exhibited no significant fluctuations in scores for responses for the same prompts ($P = 0.413$). In contrast, Kimi and Ernie Bot showed significant fluctuations ($P < 0.001$).

This indicates that, within the evaluation framework of this study, ChatGPT's output demonstrates higher internal consistency. This may be because ChatGPT was trained on open global data sources. The vast dataset enables the model to learn diverse language patterns and structures, and the wide range of data sources and adaptability helps ChatGPT maintain stable output performance.[18]

This study also shows that, despite ChatGPT's lower reliability, its consistency was superior to that of the Chinese-based models. For a LLM with high internal consistency but low reliability, greater attention should be paid to its systemic risks, and scenario modelling should be conducted. For example, for responses regarding "the best time for contraception after abortion", where ChatGPT performed worse than the other two LLMs, its use should be limited to avoid reinforcing incorrect suggestions. In standardized consultation scenarios, using high-consistency LLMs can enhance service efficiency by leveraging their consistency.

This study revealed differences in the LLMs' performance across different aspects. Specifically, Kimi and Ernie Bot's responses to questions about the best time for contraception were significantly better than those of ChatGPT (ChatGPT versus Kimi: $\chi^2$= 13.954, $P$ = 0.001; ChatGPT versus Ernie Bot: $\chi^2$= 10.366, $P$ = 0.006; Ernie Bot versus Kimi: $\chi^2$= 7.515, $P$ = 0.023). ChatGPT showed higher accuracy than the other two LLMs in responding to questions about contraceptive effectiveness (ChatGPT versus Kimi: $\chi^2$= 6.750, $P$ = 0.009; ChatGPT versus Ernie Bot: $\chi^2$= 4.320, $P$ = 0.038). For questions about potential impact of different methods of contraception on fertility, ChatGPT's responses were better than Ernie Bot's (ChatGPT versus Ernie Bot: $\chi^2$=8.043, $P$ = 0.005). For questions about the necessity of contraception after induced abortion and the choice of a contraceptive method, the responses of all three LLMs were evaluated as good. These findings suggest that different LLMs performed variably in responding to PAC-related questions. This is highly significance for selecting appropriate AI tools for specific healthcare consultation tasks. These LLMs are extremely beneficial for users seeking answers to PAC-related questions, which can help reduce the incidence of repeat abortions. Therefore, LLMs play an important role in disseminating knowledge about contraception.

This study observed a key phenomenon: LLMs significantly lag in accuracy (only 48.33% of responses were evaluated as "good"), contrasting sharply with their strengths in relevance (85.53% of responses were evaluated as "good") and completeness (75.57% of responses were evaluated as "good"). This contradiction underscores the primary risk associated with LLMs in healthcare consultations: while LLMs can generate fluent, relevant, and comprehensive responses, their reliability remains a serious concern. AI hallucinations are the primary cause of this lack of accuracy. The higher scores in other components reflect LLMs' superficial advantages in language expression and information integration. Based on the evaluation results, LLMs can be prioritized for use in PAC consultations to provide clear, highly standardized responses (such as explanations of the necessity of contraception and introduction to contraceptive methods), improving service efficiency and information coverage. However, for questions involving timing judgments (such as the best time for contraception after abortion), effect evaluations, and the impact of different contraceptive methods on fertility, LLMs' outputs must be strictly reviewed by healthcare professionals to ensure safety and reliability, as these responses are critical and prone to AI hallucinations.

When LLMs are used in PAC consultations in the Chinese context, more attention should be paid to their compatibility with traditional culture relevant to reproductive health, compliance with regulations for handling sensitive personal data, and the ethical boundaries of medical decision-making. This can ensure that the use of LLMs in PAC consultations meets the requirements of local policies.

This study has several limitations. First, it included only 20 questions from five areas based on PAC consultations, with no tests conducted for complex multi-round conversations or scenarios requiring emotional support. Therefore, the results may not be applicable to unstructured consultations. Second, the three PAC consultants were from tertiary hospitals, and their scoring criteria may not be applicable to primary healthcare institutions, potentially idealizing the LLM performance evaluation outcomes. Third, 11.1% of responses contained significantly incorrect healthcare information, unanimously identified by the PAC consultants. However, their evaluation relied on subjective judgment. Future studies should incorporate objective indicators and reliability verification to create a more robust evaluation framework. Fourth, despite independent evaluations by three PAC consultants and strict criteria, the results may only partially reflect the evaluation framework's focus on informative tasks. Future research should include rigorous validation of clinical

accuracy and methods to detect hallucinations. Fifth, this study did not simulate user questioning or correction mechanisms present in real scenarios, which may have led to an overestimation of LLMs' dynamic correction capabilities. Future studies should simulate real-world scenarios and testing conversational coherence. Sixth, the evaluation of output stability relied on one PAC consultant's scores for each LLM's three responses to each of the 20 questions. While this effectively detected statistically significant differences, it did not provide specific quantitative indicators of score fluctuations (such as standard deviation or coefficient of variation). Future studies should include these indicators to assess LLMs' output stability more comprehensively.

## Conclusions

This study was the first to investigate the performance of LLMs' responses to inquiries about PAC in the context of the Chinese language. The results showed that the three LLMs performed well overall and demonstrated significant potential for use in healthcare consultations. Chinese-based LLMs may have advantages in healthcare consultations within the Chinese language context, which is highly significant for developing AI-driven healthcare solutions tailored to specific regions. The LLMs' strengths in language expression and information integration provide an advantage in answering objective questions. However, 20 (11.1%) responses in this study were rated as "poor". AI hallucinations might be the main cause of poor accurate responses. The accuracy of LLMs' responses in clinical applications requires further improvement. Previous studies have suggested that a standardized framework should be considered for integrating LLMs into healthcare to ensure safe, effective, and equitable clinical practice.[19–21] Given that inaccurate information could have serious, potentially life-threatening impacts on patients, the accuracy of LLMs' responses must continue to be evaluated and improved in future studies.

## Data Sharing Statement

The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request.

## Ethics Approval

This study was conducted in conformity with the Declaration of Helsinki and the requirements of relevant regulations of China. Since the study did not involve the use of human or animal data, the study did not require ethics approval from an ethics committee.

## Funding

No funding was obtained for this study.

## Disclosure

The authors report no conflicts of interest in this work.

## References

1. Bridwell RE, Long B, Montrief T, Gottlieb M. Post-abortion complications: a narrative review for emergency clinicians. *West J Emerg Med*. 2022;23(6):919–925. doi:10.5811/westjem.2022.8.57929
2. Starrs AM, Ezeh AC, Barker G, et al. Accelerate progress-sexual and reproductive health and rights for all: report of the Guttmacher-Lancet Commission. *Lancet*. 2018;391(10140):2642–2692. doi:10.1016/S0140-6736(18)30293-9
3. Zhang WH, Li J, Che Y, et al. Effect of post-abortion family planning services on preventing unintended pregnancy and repeat abortion (INPAC): a cluster randomised controlled trial in 30 Chinese provinces. *Lancet*. 2017;390:S29. doi:10.1016/S0140-6736(17)33167-7
4. Temmerman M. Missed opportunities in women's health: post-abortion care. *Lancet Glob Health*. 2019;7(1):e12–e13. doi:10.1016/S2214-109X(18)30542-4
5. United Nations Report of the International Conference on Population and Development. Cairo. 1994. Available from: https://www.un.org/development/desa/pd/sites/www.un.org.development.desa.pd/files/a_conf.171_13_rev.1.pdf. Accessed January 21, 2025.
6. Zeng X, Dai H, Huang X, Xie M, Tan T. Practice effect of post-abortion care in pregnant women with heart disease. *J Nurs Sci*. 2019;34(24):1–3,14. doi:10.3870/j.issn.1001-4152.2019.24.001
7. Sallam M. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. *Healthcare*. 2023;11(6):887. doi:10.3390/healthcare11060887

8. Mondal H, De R, Mondal S, Juhi A. A large language model in solving primary healthcare issues: a potential implication for remote healthcare and medical education. *J Educ Health Promot*. 2024;13(1):362. doi:10.4103/jehp.jehp_688_23
9. Liu PR, Lu L, Zhang JY, Huo TT, Liu SX, Ye ZW. Application of artificial intelligence in medicine: an overview. *Curr Med Sci*. 2021;41 (6):1105–1115. doi:10.1007/s11596-021-2474-3
10. 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;177(2):210–220. doi:10.7326/M23-2772
11. Juexiao. Repeated abortions pose a serious risk to health and cause significant harm to the body (health internet) [Internet]. Beijing (China): People's Daily. 2015. Available from: https://paper.people.com.cn/rmrbhwb/html/2015-05/01/content_1560036.htm. Accessed 8, June 2025.
12. Zhang X, Jiang H, He Y, Liu Y. Current status and suggestions of family planning service after abortion in China. *Chin Nurs Manage*. 2019;19 (11):1720–1724. doi:10.3969/j.issn.1672-1756.2019.11.026
13. Chinese Medical Association Chinese Society of Family Planning. Guide for family planning services after an induced abortion. *Chin J Obstet Gynecol*. 2011;46(4):319–320. doi:10.3760/cma.j.issn.0529-567x.2011.04.024.
14. Cheng L, Di W, Ding Y, et al. Chinese expert consensus on the clinical application of contraceptive methods for woman. *Shanghai Med J*. 2018;41 (11):641–655.
15. Su Z, Jin K, Wu H, Luo Z, Grzybowski A, Ye J. Assessment of large language models in cataract care information provision: a quantitative comparison. *Ophthalmol Ther*. 2024;14(1):103–116. doi:10.1007/s40123-024-01066-y
16. Alkaissi H, McFarlane SI. Artificial hallucinations in ChatGPT: implications in scientific writing. *Cureus*. 2023;15(2):e35179. doi:10.7759/cureus.35179
17. Chervenak J, Lieman H, Blanco-Breindel M, Jindal S. The promise and peril of using a large language model to obtain clinical information: chatGPT performs strongly as a fertility counseling tool with limitations. *Fertil Steril*. 2023;120(3 Pt 2):575–583. doi:10.1016/j.fertnstert.2023.05.151
18. Domínguez-Díaz A, Goyanes M, de-Marcos L, Prado-Sánchez VP. Comparative analysis of automatic gender detection from names: evaluating the stability and performance of ChatGPT versus Namsor, and Gender-API. *PeerJ Comput Sci*. 2024;10:e2378. doi:10.7717/peerj-cs.2378
19. Waldock WJ, Zhang J, Guni A, Nabeel A, Darzi A, Ashrafian H. The accuracy and capability of artificial intelligence solutions in health care examinations and certificates: systematic review and meta-analysis. *J Med Internet Res*. 2024;26:e56532. doi:10.2196/56532
20. Shool S, Adimi S, Saboori Amleshi R, Bitaraf E, Golpira R, Tara M. A systematic review of large language model (LLM) evaluations in clinical medicine. *BMC Med Inform Decis Mak*. 2025;25(1):117. doi:10.1186/s12911-025-02954-4
21. Wang L, Li J, Zhuang B, et al. Accuracy of large language models when answering clinical research questions: systematic review and network meta-analysis. *J Med Internet Res*. 2025;27:e64486. doi:10.2196/64486



**Risk Management and Healthcare Policy**

**Dovepress**
Taylor & Francis Group

**Publish your work in this journal**

Risk Management and Healthcare Policy is an international, peer-reviewed, open access journal focusing on all aspects of public health, policy, and preventative measures to promote good health and improve morbidity and mortality in the population. The journal welcomes submitted papers covering original research, basic science, clinical & epidemiological studies, reviews and evaluations, guidelines, expert opinion and commentary, case reports and extended reports. The manuscript management system is completely online and includes a very quick and fair peer-review system, which is all easy to use. Visit http://www.dovepress.com/testimonials.php to read real quotes from published authors.

**Submit your manuscript here:** https://www.dovepress.com/risk-management-and-healthcare-policy-journal