Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Can Large Language Models Provide High-Quality Information for Patients with Multidrug-Resistant Organism Infections: A Multidimensional Comparative Analysis of DeepSeek-V3, ChatGPT-5.2, Gemini-3, and Grok-4

Yongxin Ma
Henan University of Chinese Medicine
Lei Li
Hennan Integrative Medicine Hospital (Hennan Academy Of Chinese Medicine)
Yige Shi
Henan University of Chinese Medicine
Zebing Ma
Hunan University of Chinese Medicine
Mengyao Liu
Henan University of Chinese Medicine
Yingying Wang
Fujian University of Chinese Medicine
Huayu Fan
Luoyang Orthopedic Hospital of Henan Province, Orthopedic Hospital of Henan Province
Xiangyang Cao
Henan Academy of Traditional Chinese Medicine Sciences
Rui Chen
41903878@qq.com

Luoyang Orthopedic Hospital of Henan Province, Orthopedic Hospital of Henan Province

Research Article

Keywords: Large language models, Multidrug-Resistant Organism Infections, Health education

Posted Date: July 1st, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9067671/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Can Large Language Models Provide High-Quality Information for Patients with Multidrug-Resistant Organism Infections: A Multidimensional Comparative Analysis of DeepSeek-V3, ChatGPT-5.2, Gemini-3, and Grok-4

Yongxin Ma[1]，Lei Li[3]，Yige Shi[1]，Zebing Ma[4]，Mengyao Liu[1]，Yingying Wang[5]，Huayu Fan[6]，Xiangyang Cao[7、8],Rui Chen[2]

1. School of Nursing（Nursing School of Healthccare Industry）, Henan University of Chinese Medicine, Henan, Zhengzhou,450000, China.

2. Department of Infection Control, Luoyang Orthopedic Hospital of Henan Province, Orthopedic Hospital of Henan Province, Zhengzhou, Henan, 450000,China.

3. Department of Preventive Treatment, Hennan Integrative Medicine Hospital (Hennan Academy Of Chinese Medicine), Zhengzhou, Henan,450000, China.

4. Hunan University of Chinese Medicine, Hunan, Changsha,410208, China

5. School of Nursing, Fujian University of Chinese Medicine, Fujain, Fuzhou,350122, China.

6. Department of Hip Injuries No.2, Luoyang Orthopedic Hospital of Henan Province, Orthopedic Hospital of Henan Province. Zhengzhou, Henan, 450000,China

7. Institute of Intelligent Medical and Bioengineering，Henan Academy of Traditional Chinese Medicine Sciences，Zhengzhou, Henan, 450000,China.

8. Henan Province Artificial Intelligence Engineering Research Center for Bone Injury Rehabilitation. Luoyang Orthopedic Hospital of Henan Province, Orthopedic Hospital of Henan Province, Zhengzhou, Henan, 450000,China.

Corresponding author：Huayu Fan, Xiangyang Cao, Rui Chen

Email ： Huayu Fan:huayulyzg@163.com Xiangyang Cao:cxy1260@126.com Rui Chen: 41903878@qq.com

Background：Large language models are being utilised with increasing frequency for patient education; however, their reliability in high-risk clinical communication remains uncertain. This study assessed the content quality, readability, and sentiment analysis of leading large language models in response to patient education queries concerning infections caused by multidrug-resistant organisms.

Methods：Ten fundamental questions were formulated following semi-structured interviews and consultations with experts from various disciplines. Multiple large language models generated responses using a standardized prompt. The quality of content was evaluated independently by seven reviewers (comprising 2 physicians, 1 nurse, and 4 infection prevention and control professionals) using 5-point Likert scales across five domains: accuracy, comprehensiveness, safety, humanistic care, and practicality. Variations between groups were analysed using SPSS. Readability was determined in R through the Flesch–Kincaid Grade Level, Gunning Fog Index, and Flesch Reading Ease. Additionally, sentiment analysis was performed in R.

Results：Expert ratings varied significantly across models in all five domains (all $p < 0.001$). Grok-4 demonstrated the highest overall performance, achieving top scores in accuracy (4.13 ± 0.64), comprehensiveness (4.31 ± 0.62), safety (4.16 ± 0.58), humanistic care (3.51 ± 0.82), and practicality (4.09 ± 0.84). In contrast, ChatGPT-5.2 received the lowest ratings across several domains. Readability analyses revealed that all model outputs surpassed the American Medical Association's recommended sixth-grade reading level for patient materials. Sentiment analysis indicated a positive tendency for DeepSeek-V3 (0.064), neutrality for ChatGPT-5.2 (0.013), and negative tendencies for Grok-4 (-0.051) and Gemini-3 (-0.087).

Conclusions：In the context of patient education for multidrug-resistant organisms, a high-stakes environment, no single large language model has achieved acceptable standards in terms of quality, safety, humanistic care, and practicality. Consequently, large language models should be employed solely as tools for assisted drafting, necessitating mandatory expert review and contextual adaptation to guarantee scientific rigor, patient safety, and comprehensibility within a careful regulatory framework.

Keywords：Large language models; Multidrug-Resistant Organism Infections; Health education;

1 Introduction

The extensive use of antibacterial drugs has led to a rise in infections caused by

multidrug-resistant organisms (MDRO), posing a significant threat to public health and medical safety[1]. MDRO infections not only complicate treatment and increase medical costs but also extend hospital stays, elevate the risk of complications and mortality[2,3]. Moreover, the persistent spread of MDRO within hospital settings escalates the challenges and prerequisites for infection prevention and control[4]。 Given the substantial disease burden associated with MDRO infections and their transmission risks in hospitals, infection control measures cannot solely rely on healthcare professionals but necessitate active involvement of patients and their families. They play a crucial role in implementing fundamental practices like isolation and hand hygiene[5]. Improving their understanding of drug resistance and transmission risks is pivotal not only for enhancing the healthcare experience and satisfaction but also for fostering better clinical outcomes[6,7].

However, in clinical Settings, there is a significant cognitive gap and behavioral discontinuity in the health education of patients and their families infected with MDRO. Most patients and their families remain aware of the presence of drug-resistant bacteria but find it difficult to understand their transmission routes, key points of prevention and control, and the necessity of this[8]. On the one hand, information related to multi-drug resistant bacteria involves professional terms and concepts of infection prevention and control. When medical staff do not explain it adequately, patients often show insufficient understanding and uncertainty, making it difficult for them to form a correct understanding of diagnosis and prevention and control measures in a short period of time[9]； On the other hand, the clinical work is fast-paced and human resources are limited. In practice, ward education mainly relies on time-consuming and laborious oral education[10]， The pressure of resources and time also leads to insufficient communication and educational gaps regarding the risks of antibacterial drugs, thereby affecting patients' understanding and compliance with prevention and control measures[11]。

Meanwhile, Large Language Models (LLMs) are advancing rapidly, and their potential applications in medical consultation and health education are becoming increasingly significant[12,13]. An increasing number of patients are inclined to utilise

LLMs to acquire health information, including details about their health status, medication, and disease diagnosis[14,15]。In the context of multi-drug resistant organism (MDRO) infections, patients often encounter isolation signs, visiting restrictions, or medical protective equipment, which heightens their need for explanations and emotional support, prompting them to seek additional information sources to alleviate uncertainty[16]. However, LLMs occasionally demonstrate a phenomenon termed "hallucinations," resulting in the generation of inaccurate or misleading responses, which considerably undermines the model's credibility[17]. Particularly concerning the high-risk subject of MDRO infections, which encompasses nosocomial transmission, isolation strategies, and antibiotic management, the output from LLMs must be both scientifically rigorous and comprehensible. Nevertheless, the reliability of the responses generated by large language models to patients' real-world inquiries in the critical area of infection prevention and control remains uncertain at present.

This study addresses the core concerns of patients and their families in real-life scenarios by evaluating the performance of DeepSeek V3, ChatGPT-5.2, Gemini-3, and Grok-4 in producing educational materials for MDRO patients. The assessment focuses on dimensions such as accuracy, comprehensiveness, safety, humanistic care, and clinical practicality. By considering readability indicators and emotional expression tendencies, we compare these models to illustrate their communication styles, strengths, limitations, and associated risks in educating individuals about infection prevention and control. This analysis offers empirical support for the cautious integration, expert assessment, and standardised use of Language Model Models in educating MDRO patients and promoting infection prevention and control.

2 Methods

2.1 Study Design

As this study did not involve human subjects, ethical approval was not required. All procedures were, however, conducted in strict adherence to the principles of the Declaration of Helsinki and local ethical guidelines.

The overall study process is illustrated in Figure 1. Through semi-structured

interviews with patients suffering from MDRO and their families regarding their experiences with illness and caregiving, we summarised the core questions that elicited the greatest concern and were frequently posed. These questions were closely related to the implementation of infection prevention and control measures. A multidisciplinary team of experts was convened to conduct a structured consultation and revision process. Consequently, the question bank was screened, merged, and optimised in terms of wording, leading to the identification of ten core questions that encompass the essential areas of MDRO patient education: basic concepts, modes of transmission, susceptible populations, causes of infection, impacts on patients, treatment and cooperation, prognosis and long-term effects, caregiver self-protection, and support for comfort and rehabilitation during caregiving, as well as preventive measures (Additional file1).

Secondly, elicit responses from the LLMs following the standardized instruction template. Provide four LLM responses to each question using the prompt word: "Please respond to the following inquiries regarding multi-drug resistant bacterial infections for patients and their families affected by such infections." Ensure that the output data is well-structured and suitable for direct transfer to Word without any formatting issues. All standardized simulations were conducted by a sole author (MYX) using identical equipment, web browser, and Internet connection. The outcomes were acquired via freshly created accounts. Each question was generated independently in a separate dialogue box, and the original output text was preserved for analytical purposes.

The four models were encoded anonymously as LLM-A, LLM-B, LLM-C, and LLM-D (A: DeepSeek V3, B: ChatGPT-5.2, C: Gemini-3, D: Grok-4). All identifying information regarding the models was removed prior to the submission of the evaluation. The responses to the same question were presented in a random sequence. The project leader possesses a thorough understanding of the relationship between the coding and the sequence. Data collators exclusively managed anonymous text. Statistical analysts conducted the analysis without any blind intervention throughout the entire process.

2.2 Content Quality Evaluation

An evaluation team comprising two clinical doctors, one nurse, and four infection prevention and control Professional from three prestigious hospitals undertook a comprehensive assessment of the model output's content quality. All evaluators possessed a master's degree or higher and had between 10 and 25 years of professional experience (Additional file 2). Following standardised training, the evaluators independently rated all texts using a Likert scale ranging from 1 to 5. The scoring dimensions included accuracy, comprehensiveness, safety, humanistic care, and clinical practicality. (The evaluation criteria are elaborated in Additional file 3.)

2.3 Readability evaluation

Readability is defined as the level of reading proficiency necessary for comprehending written text[18,19].The readability of each response across the four models was quantitatively assessed using the Flesch-Kincaid Grade Level (FKGL), Gunning Fog Index (GFI), and Flesch Reading Ease Score (FRES). FKGL evaluates the grade level required for text comprehension based on sentence length and syllable count; a higher value indicates a greater educational attainment necessary for understanding[20].GFI estimates the years of education needed to comprehend a particular text, serving as a measure of the complexity of both text structure and vocabulary. It is widely accepted that health education materials intended for the general public should ideally target a reading level around grade 8, whereas texts requiring comprehension above grade 17 are typically suited for graduate-level readers[21].FRES measures reading difficulty, emphasising the number of words per sentence and the syllable count of individual words. A higher score indicates easier readability and a reduced cognitive load for comprehension.

2.4 sentiment analysis

Lexicographical sentiment analysis was employed, utilising the sentimentr package to evaluate the sentiment of health education texts aimed at patients with multi-drug resistant bacteria, which were generated by four large language models. The sentimentr package identifies positive and negative words at the sentence level, adjusting the emotional direction and intensity in conjunction with valence shifters

within the contextual framework. Thereafter, the sentence scores can be aggregated to yield ave_sentiment. The emotional score is represented as a continuous value, where negative values denote negativity, positive values signify positivity, and a score of 0 indicates neutrality. The texts were categorised based on ave_sentiment as follows: positive (≥ 0.05), neutral (-0.05 to 0.05), and negative (< 0.05).

2.5 Statistical Analysis

The expert scoring data were subjected to statistical analysis using SPSS 27.0. To assess the consistency of scores assigned by eight experts, the two-way random effects model was employed, with absolute agreement defining consistency. The intraclass correlation coefficient (ICC) and its 95% confidence interval were calculated. Given that the research objective is to evaluate the overall consistency of scores provided by multiple experts, this study reports the average measures ICC. Continuous variables with normal and non-normal distributions were reported as mean±standard deviation and median (interquartile range), respectively. Each expert evaluated four large language models across various dimensions, resulting in a repeated measurement (paired) design. The Friedman test was utilised to compare the overall differences in scores among the four models within each dimension. If the results of the Friedman test were statistically significant, the Wilcoxon signed-rank test was subsequently employed for post hoc pairwise comparisons, with the Bonferroni method applied to adjust the significance levels for multiple comparisons. After correction, α' was set at 0.05/6 = 0.0083. Categorical variables were represented by frequency. With the exception of the pairwise comparisons conducted after applying the corrected significance level, all other tests were two-sided, with $p < 0.05$ considered statistically significant.

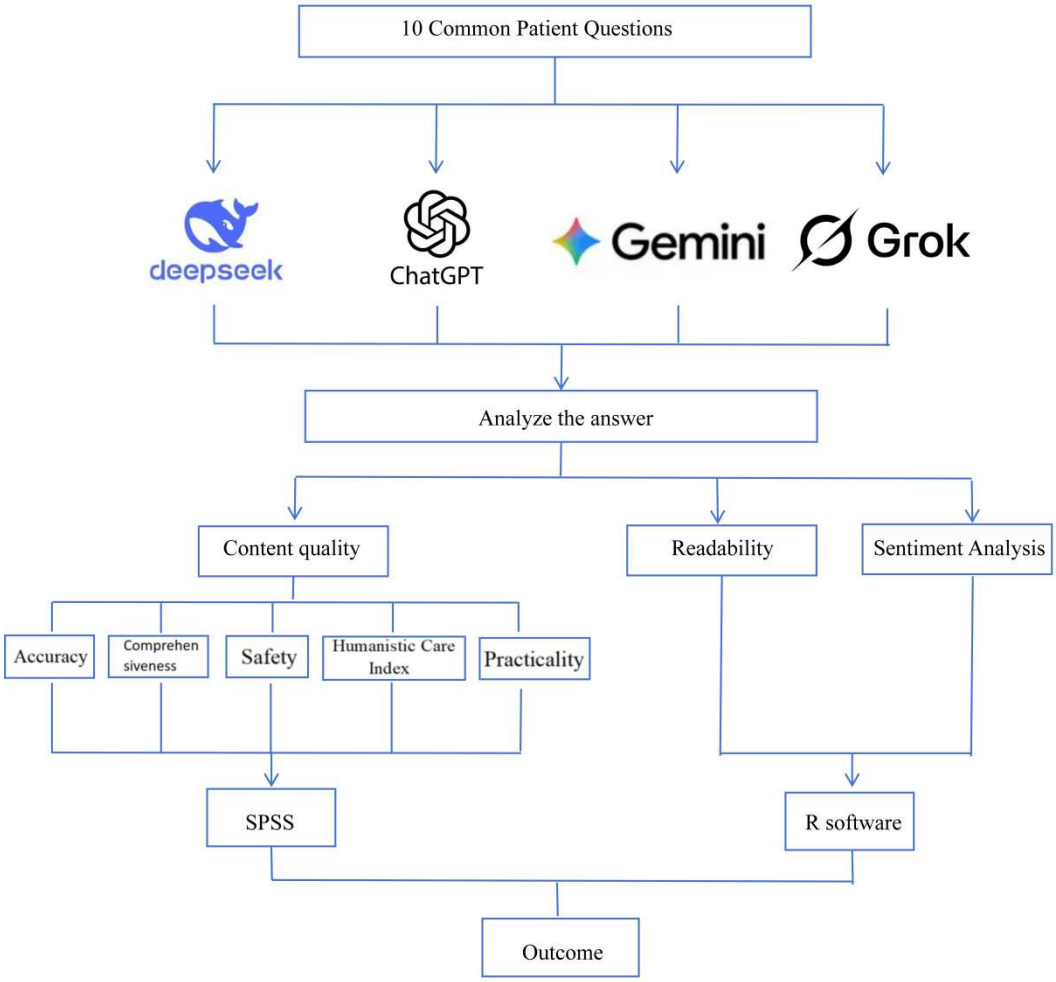


Figure 1 Overall research design roadmap

3 Results

3.1 Accuracy, Comprehensiveness, Safety, Humanistic Care and Practicality Analysis

The rater reliability analysis indicated that the scores from the seven experts exhibited good consistency across five dimensions: the mean measurement ICC (two-way random effects, absolute agreement) was 0.758, with a 95% confidence interval of 0.687-0.811, and a p-value of less than 0.001, suggesting that the scoring results are both stable and reliable. The Friedman test was employed to assess the overall differences among four large language models across the five dimensions of accuracy, comprehensiveness, security, humanistic care, and practicality.（Figure 2） The findings revealed that the overall differences among the models in each dimension were statistically significant (all $p < 0.001$). Additionally, the Wilcoxon

symbolic rank test was utilised for post hoc pairwise comparisons, with the Bonferroni method applied to adjust the significance level ($\alpha' = 0.05/6 = 0.0083$, two-tailed).

Following a normality test of the continuous variable, it was determined that the data did not conform to a normal distribution; thus, it is represented by the median M (P_{25}, P_{75}).（Table 1）

Accuracy: Post hoc comparisons revealed that Grok-4 score was significantly higher than that of ChatGPT-5.2 ($Z = -5.228$, $p < 0.001$) and Gemini-3 ($Z = -3.860$, $p < 0.001$).

Comprehensiveness: Post hoc comparisons indicated that the scores of Grok-4, DeepSeek-V3, and Gemini-3 were all significantly higher than those of ChatGPT-5.2 (all $p < 0.0083$).

Security: Post hoc comparisons demonstrated that the scores of DeepSeek-V3 and Grok-4 were significantly higher than those of ChatGPT-5.2 (DeepSeek V3 vs ChatGPT-5.2: $Z = -3.530$, $p < 0.001$; Grok-4 vs ChatGPT-5.2: $Z = -3.846$, $p < 0.001$). No statistically significant differences were observed among the remaining models ($p \geqslant 0.0083$).

Humanistic care: Post hoc comparisons revealed that the scores of Grok-4, DeepSeeking V3, and Gemini-3 were all significantly higher than those of ChatGPT-5.2 ($p < 0.0083$).

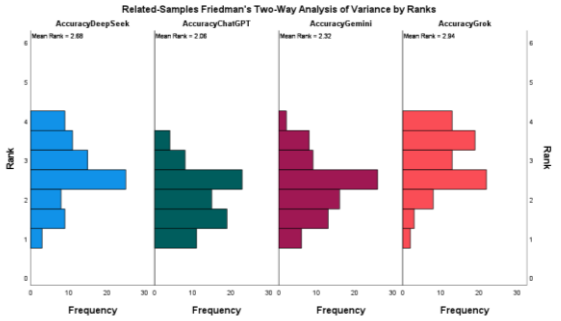
Practicality: Post hoc comparisons indicated that the scores of Grok-4 and DeepSeek V3 were significantly higher than those of ChatGPT-5.2 and Gemini-3 ($p < 0.0083$).

Table 1 shows the content quality scoring results of four large language models

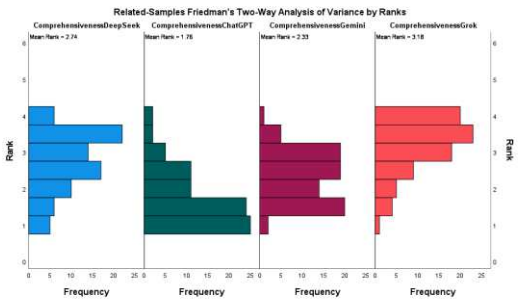
	Accuracy	Comprehensiveness	Security	Humanistic Care Index	practicality
Deepseek	4.0（4.0，4.0）	4.0（4.0，4.0）	4.0（4.0，4.0）	4.0（3.0，4.0）	4.0（3.0，4.0）
ChatGPT	4.0（3.0，4.0）	3.0（3.0，4.0）	4.0（4.0，4.0）	3.0（2.0，4.0）	3.0（2.0，4.0）
Gemini	4.0（3.0，4.0）	4.0（3.0，4.0）	4.0（4.0，4.0）	3.0（3.0，4.0）	3.5（3.0，4.0）
Grok	4.0（4.0，5.0）	4.0（4.0，5.0）	4.0（4.0，4.75）	4.0（3.0，4.0）	4.0（3.0，5.0）

Figure 2 The Friedman test results of content quality scores for four large language model：（1）Accuracy （2）Comprehensiveness （3） safety（4）humanistic care （5）practicality

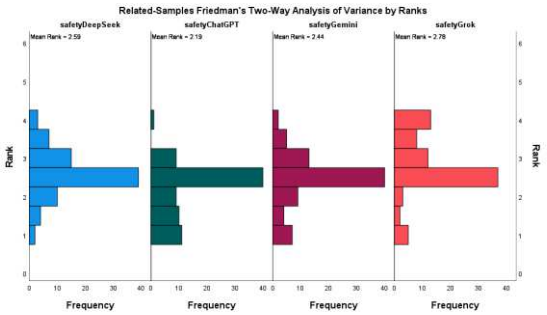
(1)



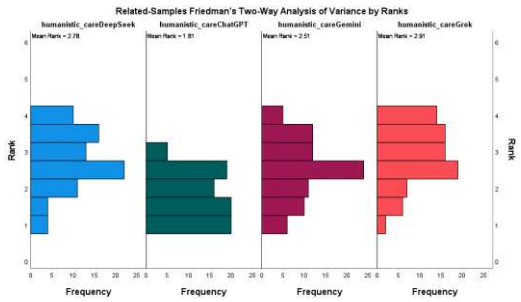
(2)



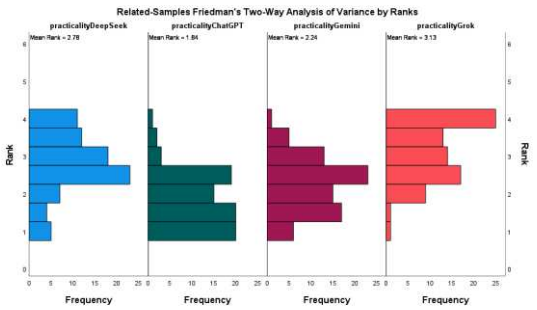
(3)



(4)



(5)



3.2 Readability Analysis

Readability analysis shows that there are significant differences among the outputs of the four models in three readability metrics (as shown in Figure 3). Specifically(Table 2), In terms of Flesch Reading Eas, Gemili-3 had the highes

t average score (45.46±13.23), followed by DeepSeek V3 (39.80±6.98) and G rok-4 (32.24±7.08). ChatGPT-5.2 had the lowest score (18.92±6.29), suggestin g that its text was overall more difficult to read. Consistent with this, in terms of Flesch-Kincaid Grade, Gemini-3 (10.60±2.07) and DeepSeek V3 (11.02±1. 21) were lower, while Grok-4 was in the middle (11.94±1.31). However, Chat GPT-5.2 had the highest level (15.72±1.28), suggesting that it requires a highe r level of education to understand. The Gunning Fog index also showed the sa me trend: Gemi-3 was the lowest (14.00±2.29), followed by DeepSeek V3 (14. 95±1.64), Grok-4 was slightly higher (15.68±1.50), and ChatGPT-5.2 was the highest (20.13±1.85).

Table 2 Readability Scoring results of Four large language models

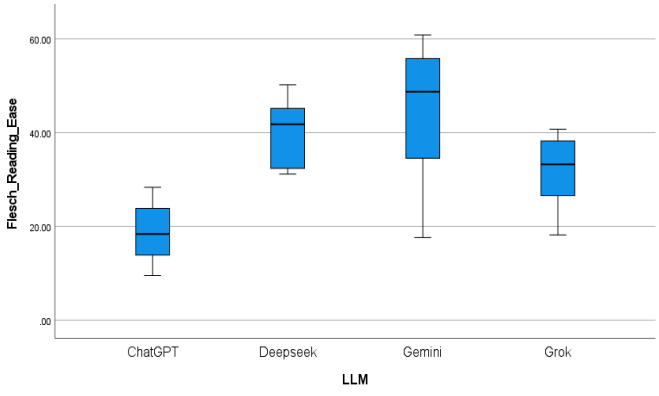
	Avg_Sentence_Length (Mean±SD)	Avg_Syllables_per_Word (Mean±SD)	Flesch Reading Ease (Mean±SD)	Flesch－Kincaid Grade (Mean±SD)	Gunning Fog Index (Mean±SD)
Deepseek-V3	13.34±2.34	1.81±0.08	39.80±6.98	11.02±1.21	14.95±1.64
ChatGPT-5.2	20.66±2.70	1.97±0.07	18.92±6.29	15.72±1.28	20.13±1.85
Gemini-3	14.83±2.24	1.72±0.15	45.4±13.23	10.60±2.07	14.00±2.29
Grok-4	12.83±2.02	1.90±0.07	32.24±7.08	11.94±1.31	15.68±1.50

a: Both Flesch-kincaid Grade and Gunning Fog Index indicate the educational attainment required to understand the text; The lower the values of these indicators are, the better the readability will be.

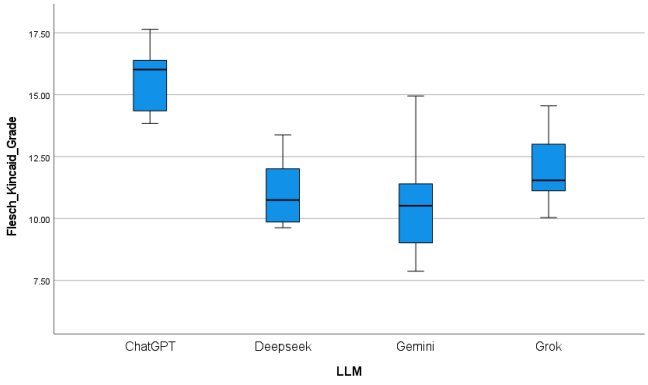
b: The higher the Flesch Reading Ease score is, the more friendly the text is and the less difficult it is to read.

Figure 3 Four large language models generate text readability test results （1）Flesch Reading Eas score（2）Flesch－Kincaid Grade Level （3）Gunning Fog Index

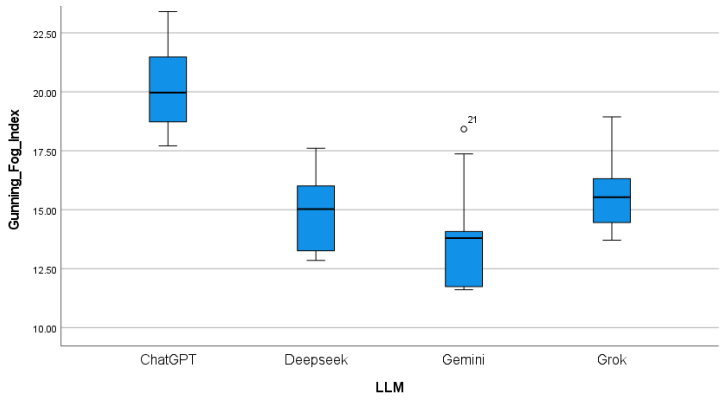
（1）



（2）



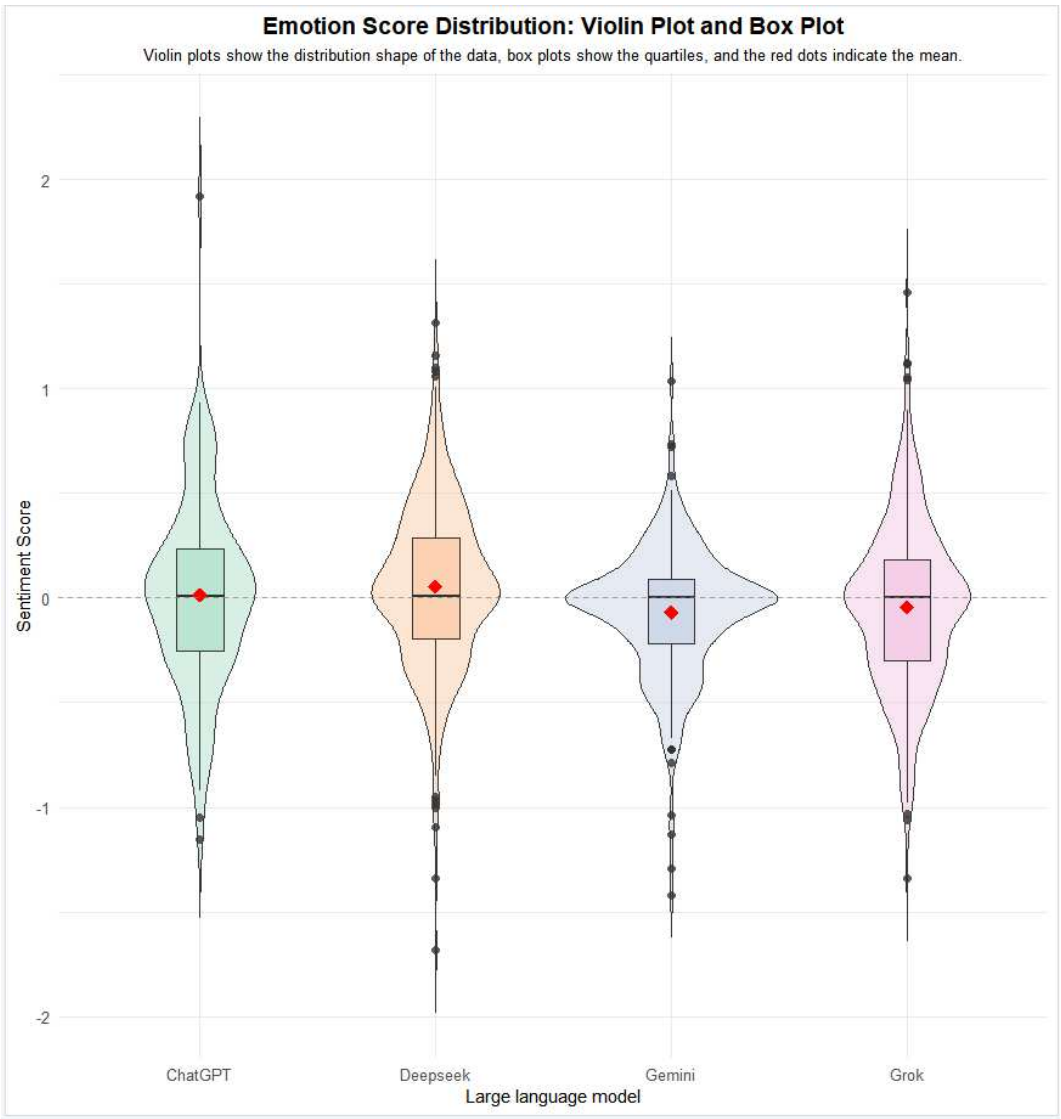
（3）



3.3 Sentiment analysis

As shown in Figure 4, the average sentiment scores of texts generated by different large language models vary, demonstrating different sentiment tendencies. Deepseek shows the strongest positive emotional tendency. Its average emotional score was 0.064, the highest among all models, and it was classified as a positive category. This indicates that the health education materials generated by Deepseek have the most optimistic or encouraging tone, but they are widely distributed and have certain emotional fluctuations. ChatGPT shows a neutral emotional tendency. Its average emotional score is 0.013, close to 0, and it is classified as a neutral category. The distribution is relatively compact, showing high consistency, and the occurrence frequency of extreme emotional values is low. This reflects that when ChatGPT generates such medical texts, it tends to maintain an objective and calm tone. Both Grok and Gemini show negative emotional tendencies. Among them, Grok's average score is -0.051, while Gemini's average score is -0.087. Both are classified as negative categories, and Gemini is the most negative, with the main part of its distribution below 0. This might imply that the texts generated by these two models place more emphasis on highlighting the severity or risk of the disease, thereby leading to a lower emotional score.

Figure 4 The sentiment analysis results of four large language models



4 Discussion

This study systematically evaluated the performance of four types of LLMs in generating patient education materials based on actual inquiries from patients with MDRO and their families. The evaluation encompassed expert assessments of content quality, including accuracy, comprehensiveness, safety, humanistic care, and clinical practicality, alongside readability metrics and sentiment analysis. Overall, all four models produced relatively complete and well-structured educational texts addressing core issues, indicating that LLMs possess a viable foundation for health education concerning infection prevention and control. However, notable disparities exist among the models regarding content quality, reading difficulty, and emotional communication style. These differences may be exacerbated in the context of MDRO,

which is characterised by a high transmission risk, stringent isolation protocols, and sensitivity to antibiotic management. Therefore, thorough assessment and control measures must be implemented prior to clinical adoption.

4.1 Content Quality

This study performed a multi-dimensional comparison of the quality of patient education texts produced by four LLM within the context of high-risk infection prevention and control concerning MDROs. The findings revealed significant disparities among the models regarding accuracy, comprehensiveness, safety, humanistic care, and practicality. Effective health education on MDROs hinges not only on the accuracy of the information provided but also on its ability to encompass the critical aspects of executable prevention and control while clearly delineating behavioural boundaries; thus, comprehensiveness and practicality are essential. LLMs do not function universally; superior capabilities do not necessarily equate to broader applicability[22]，In the realm of health education focused on infection prevention and control, it is imperative to achieve a balance among accurate, comprehensive, practical, and safe knowledge.

It is noteworthy that the scores for all models regarding humanistic care are generally low. Even those models demonstrating relatively superior performance achieve scores that are only marginally above the median level. Isolation measures associated with MDRO infections frequently lead patients to feel restricted, treated differently, and even stigmatized, resulting in negative emotions such as isolation, anxiety, and fear[23], Consequently, high-quality patient education should not solely concentrate on the transmission of factual information; it must also alleviate patients' psychological burdens and enhance their willingness to cooperate through the provision of humanistic care[24,25], While the current text generated by LLMs can deliver structured information, it often lacks emotional buffer statements and patient-centred communication strategies. In the realm of infection prevention and control, such a communication style may not only diminish patients' acceptance of isolation measures but also undermine the foundation of trust between doctors and patients, as well as the process of joint decision-making.

4.2 Readability

The analysis of readability results exposes a discrepancy between the coherence of information and the intricacy of language in the text produced by the model. While certain models achieve high scores in content quality, their readability levels far surpass the advised benchmarks for public health education. Guidelines from the American Medical Association (AMA) and the National Institutes of Health (NIH) stipulate that the readability of public health educational materials should not surpass the sixth-grade level[26]. Nevertheless, even the most comprehensible model examined in this research surpassed this criterion.

This finding aligns with prior research on the information generation capabilities of LLMs in various medical domains[27,28]. The underlying mechanism may stem from the disproportionately high prevalence of professional texts within the model's training corpus, which encourages the model to maintain the integrity of terminology when elucidating medical concepts[29]. In the context of MDROs, this inclination is further exacerbated; concepts such as drug resistance mechanisms, contact transmission, and isolation protection are inherently abstract and interconnected, necessitating complex behavioural decisions. In striving for rigor and completeness, models frequently employ longer sentence structures and a denser array of terms, thereby complicating readability.

From a clinical application standpoint, readability encompasses not only the extent to which information can be understood but also its direct impact on practical executability. For individuals with low health literacy, an excessively high reading threshold may result in comprehension bias, reduced adherence, and inadequate implementation of family protective behaviours[30]. Consequently, when integrating LLMs for educational purposes within clinical practice, readability should be considered an evaluative dimension of equal significance to content quality.

4.3 Sentiment Analysis

Sentiment analysis elucidates the implicit variations in information transmission across different models. The average sentiment score of DeepSeek V3 indicates a positive inclination, suggesting that its text is more encouraging or supportive. This

attribute is particularly significant in the context of infection prevention and control education. The expression of positive emotions enhances patients' acceptance and self-efficacy, particularly in discharge guidance, family care advice, and scenarios where anxiety reduction is essential; it can bolster patients' confidence and coping abilities[31]. In contrast, the language styles of Grok-4 and Gemini-3 exhibit a slightly negative tone, focusing on potential risks and the requirements for prevention and control.

In the context of MDRO, patients frequently encounter feelings of anxiety, shame, or stigmatization upon observing isolation signage, medical protective equipment, and visitation restrictions[32,33]. The use of excessively negative language may exacerbate perceptions of threat and stigma, resulting in avoidance of communication or resistance to isolation measures[34], Conversely, overly positive messaging may trivialise the associated risks and diminish the focus on infection prevention behaviours[35]. Therefore, optimal patient education materials must achieve a balance between conveying risk warnings and instilling hope.

Various models may be relevant to distinct stages of communication. For example, in the initial diagnosis and isolation notification phase, a moderate emphasis on risks could facilitate behavioural adjustment. Conversely, during the discharge guidance or long-term management phase, a supportive tone may better enhance self-efficacy. Future research could investigate model selection strategies grounded in scenario stratification to offer more personalised and contextualised emotional support for patient education across different stages.

4.4 Risk and Ethical Issues

The information pertaining to MDRO exhibits clear scene dependence, necessitating that intervention measures be closely aligned with the specific medical environment[4]. Additionally, due to the disparities in policies and resource availability across various hospitals, significant variations in prevention and control strategies often arise among different institutions[36].If the content generated by the model fails to explicitly underscore the necessity of adhering to our hospital's protocols, patients may erroneously perceive it as a universal standard, potentially leading to the

adoption of inappropriate prevention and control measures. Moreover, if the terminology regarding antibacterial drugs is susceptible to misinterpretation, it may prompt patients to engage in improper self-medication practices[37], thereby exacerbating the issue of drug resistance and posing an increased public health risk.

The appropriate role of LLMs in patient education should be as an information assistance tool rather than a substitute for decision-making[38]. While structured educational content can be delivered efficiently, it remains essential for professionals to review and modify this content in practical applications to guarantee its accuracy and alignment with the specific requirements of clinical contexts[39,40].

4.5 Practical Implications

This study establishes a three-tiered application framework aimed at standardising the utilisation of LLMs in health education for infection prevention and control. The first tier, initial draft generation, employs the model to efficiently produce structured educational texts, thereby enhancing the effectiveness of preliminary content creation. The second tier, professional proofreading, entails verification of key information by experts in infection prevention and control, ensuring both accuracy and adherence to relevant policies, which guarantees the content's rigor and reliability. The third tier focuses on language and emotional optimisation. By simplifying expressions, incorporating empathetic language, and adjusting tonal intensity, the readability and humanistic quality of the text are improved. Furthermore, model developers should implement specialised optimisations for patient education contexts, which include enhancing the training of accessible corpora, introducing mechanisms for readability control, and incorporating risk warning messages at the output stage.

4.6 Limitations

In this study, we employed uniform cue words to ensure comparability; however, the sensitivity of various models to cue strategies may differ. Future research could further investigate the moderating effect of cue engineering on output quality. Additionally, the number of questions assessed in this study is limited. While it encompasses the core areas of MDRO patient education, it does not address the

complex scenarios encountered across different departments and types of resistant bacteria. Lastly, despite the consistency of expert scores, subjective factors remain. Future evaluations could incorporate patient participation to more accurately reflect the understanding and acceptance of the target population.

5 Conclusions

The utilisation of LLMs in health education for infection prevention and control necessitates caution at this time. In clinical practice, LLMs may function as supplementary tools for generating educational materials, thereby enhancing efficiency and the breadth of information provided; however, they should not supplant professional proofreading and contextual adjustments. Future research should investigate strategies for optimising readability, enhancing emotional communication mechanisms, and developing a model selection framework grounded in scenario stratification. Additionally, it is essential to establish a continuous evaluation and oversight mechanism to ensure the judicious application of LLMs in the domain of infection prevention and control, while upholding safety and ethical standards.

List of abbreviations

MDRO:Multidrug-resistant organism

LLMs:Large Language Models

FKGL:Flesch-Kincaid Grade Level

GFI:Gunning Fog Index

FRES:Flesch Reading Ease Score

ICC:intraclass correlation coefficient

Declarations

Ethics approval and consent to participate

The study strictly followed the Declaration of Helsinki, and ethic approval was obtained from the Ethics Committee of Luoyang Orthopedic Hospital of Henan Province, Orthopedic Hospital of Henan Province.(No. 2025KYKT0047-01).Written informed consent was obtained from all experts involved in the scoring and evaluation process prior to their participation in the study.

Consent for publication

Not Applicable

Availability of data and materials

All data generated or analysed during this study are included in this published article [and its supplementary information files].

Competing interests

The authors declare that they have no competing interests

Funding

No funds, grants, or other support were received.

Authors' contributions

MYX:manuscript drafting or manuscript revision for important content, data curation, experimental design, experimental supervision, and thesis modification;:LL and SYG:Expert Rating Data Collection and Analysis; MZB and LMY:Data analysis and paper revision;WYY:Data collection and organization;FHY and CXY:critical revision of the manuscript for important intellectual content, methodological and statistical guidance, experimental supervision; CR: experimental design, experimental supervision, methodology, manuscript drafting, and manuscript modification.All authors read and approved the final manuscript.

Acknowledgements

Not applicable

References

[1] Liew K C, O'Keeffe J, Rajandas H, et al. Insights into the evolution of P. aeruginosa antimicrobial resistance in a patient undergoing intensive therapy[J]. Antibiotics, 2023, 12(3): 483.

[2] Nasser S, Alnasser Z, Aljuhani O, et al. Exploring infection risk factors and multi-drug-resistant organisms (MDROs) in burn intensive care units: a multi-centre case-control study[J]. Journal of Hospital Infection, 2025, 162: 186-196.

[3] Nelson R E, Hyun D, Jezek A, et al. Mortality, length of stay, and healthcare costs associated with multidrug-resistant bacterial infections among elderly hospitalized patients in the United States[J]. Clinical Infectious Diseases: an Official Publication of the Infectious Diseases Society of America, 2022, 74(6): 1070-1080.

[4] Geng Y, Liu Z, Ma X, et al. Infection prevention and control measures for multidrug-resistant organisms: a systematic review and network meta-analysis[J]. Infection, 2025, 53(5): 1789-1800.

[5] Donskey C J. Empowering patients to prevent healthcare-associated infections[J]. American Journal of Infection Control, 2023, 51(11S): A107-A113.
[6] Gillitzer V, Rath A, Caplunik-Pratsch A, et al. Counselling sessions for patients in contact isolation due to multi-drug-resistant organisms improve informedness and reduce dissatisfaction[J]. Journal of Hospital Infection, 2024, 154: 53-59.
[7] Gaube S, Däumling S, Biebl I, et al. Patients with multi-drug-resistant organisms feel inadequately informed about their status: adverse effects of contact isolation[J]. Journal of Hospital Infection, 2023, 133: 89-94.
[8] Watson E, Tsindos T, Peleg A Y, et al. Understanding patient and healthcare worker experiences and perspectives of multidrug-resistant organisms[J]. JAC-Antimicrobial Resistance, 2023, 5(3): dlad071.
[9] Wojcik G, Afseth J, Fagan R, et al. Patient and public understanding of antimicrobial resistance: a systematic review and meta-ethnography[J]. JAC-Antimicrobial Resistance, 2024, 6(4): dlae117.
[10] Yang L, Qin Z, Yao J, et al. Enhancing patient satisfaction and reducing nurse workload: the impact of multimedia health education in a prospective single-center randomized controlled trial[J]. Frontiers in Medicine, 2025, 12: 1400061.
[11] Balea L B, Gulestø R J A, Xu H, et al. Physicians', pharmacists', and nurses' education of patients about antibiotic use and antimicrobial resistance in primary care settings: a qualitative systematic literature review[J]. Frontiers in Antibiotics, 2024, 3: 1507868.
[12] Ayers J W, Poliak A, Dredze M, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum[J]. JAMA Internal Medicine, 2023, 183(6): 589-596.
[13] You Q, Luo Y, Chen Z, et al. How large language model responds to common depression questions: a comparative analysis of ChatGPT-4.0, DeepSeek, google gemini and perplexity[J]. Nurse Education in Practice, 2026, 90: 104668.
[14] Mendel T, Singh N, Mann D M, et al. Laypeople's use of and attitudes toward large language models and search engines for health queries: survey study[J]. Journal of Medical Internet Research, 2025, 27: e64290.
[15] Yun H S, Bickmore T. Online health information-seeking in the era of large language models: cross-sectional web-based survey study[J]. Journal of Medical Internet Research, 2025, 27: e68560.
[16] Baron R, Eilers R, Haverkate M R, et al. A qualitative study examining the impact of multidrug-resistant organism (MDRO) carriage on the daily lives of carriers and parents of carriers with experiences of hospital precautionary measures[J]. Antimicrobial Resistance & Infection Control, 2022, 11(1): 103.
[17] Safranek C W, Sidamon-Eristoff A E, Gilson A, et al. The role of large language models in medical education: applications and implications[J]. JMIR Medical Education, 2023, 9: e50945.
[18] Roberts H, Zhang D, Dyer G S M. The readability of AAOS patient education materials: evaluating the progress since 2008[J]. Journal of Bone and Joint Surgery. American Volume, 2016, 98(17): e70.
[19] Friedman D B, Hoffman-Goetz L. A systematic review of readability and comprehension instruments used for print and web-based cancer information[J]. HEALTH Education &

Behavior: the Official Publication of the Society for Public HEALTH Education, 2006, 33(3): 352-373.
[20] Flesch R. A new readability yardstick[J]. Journal of Applied Psychology, 1948, 32(3): 221-233.
[21] Lv X, Zhang X, Li Y, et al. Leveraging large language models for improved patient access and self-management: assessor-blinded comparison between expert- and AI-generated content[J]. Journal of Medical Internet Research, 2024, 26: e55847.
[22] Shool S, Adimi S, Saboori Amleshi R, et al. A systematic review of large language model (LLM) evaluations in clinical medicine[J]. BMC Medical Informatics and Decision Making, 2025, 25(1): 117.
[23] Wijnakker R, Lambregts M M C, Rump B, et al. Limited multi-drug resistant organism related stigma in carriers exposed to isolation precautions: an exploratory quantitative questionnaire study[J]. Journal of Hospital Infection, 2020, 106(1): 126-133.
[24] Zemni I, Gara A, Nasraoui H, et al. The effectiveness of a health education intervention to reduce anxiety in quarantined COVID-19 patients: a randomized controlled trial[J]. BMC Public Health, 2023, 23(1): 1188.
[25] Guan T, Qan'ir Y, Song L. Systematic review of illness uncertainty management interventions for cancer patients and their family caregivers[J]. Supportive Care in Cancer: Official Journal of the Multinational Association of Supportive Care in Cancer, 2021, 29(8): 4623-4640.
[26] Bange M, Huh E, Novin S A, et al. Readability of patient education materials from RadiologyInfo.org: has there been progress over the past 5 years?[J]. AJR. American Journal of Roentgenology, 2019, 213(4): 875-879.
[27] Shukla I Y, Sun M Z. Online and ChatGPT-generated patient education materials regarding brain tumor prognosis fail to meet readability standards[J]. Journal of Clinical Neuroscience, 2025, 138: 111410.
[28] Fahy S, Oehme S, Milinkovic D, et al. Assessment of quality and readability of information provided by ChatGPT in relation to anterior cruciate ligament injury[J]. Journal of Personalized Medicine, 2024, 14(1): 104.
[29] Gu Y, Tinn R, Cheng H, et al. Domain-specific language model pretraining for biomedical natural language processing[J]. ACM Transactions on Computing for Healthcare, 2022, 3(1): 1-23.
[30] Miller T A. Health literacy and adherence to medical treatment in chronic and acute illness: a meta-analysis[J]. Patient Education and Counseling, 2016, 99(7): 1079-1086.
[31] Heyn L G, Løkkeberg S T, Ellington L, et al. Understanding the role of positive emotions in healthcare communication – a realist review[J]. Nursing OPEN, 2023, 10(6): 3447-3459.
[32] Baubie K, Shaughnessy C, Safdar N. Exploring patient perceptions of contact precautions[J]. American Journal of Infection Control, 2019, 47(2): 225-226.
[33] Gammon J, Hunt J, Musselwhite C. The stigmatisation of source isolation: a literature review[J]. Journal of Research in Nursing: JRN, 2019, 24(8): 677-693.
[34] Volkow N D, Gordon J A, Koob G F. Choosing appropriate language to reduce the stigma around mental illness and substance use disorders[J]. Neuropsychopharmacology, 2021, 46(13): 2230-2232.
[35] Park T, Ju I, Ohs J E, et al. Optimistic bias and preventive behavioral engagement in the context of COVID-19[J]. Research in Social and Administrative Pharmacy, 2021, 17(1):

1859-1866.
[36] Coffey K C, Grossman T, Banach D B, et al. Current infection control practices for multidrug-resistant organisms (MDRO): a survey of the society for healthcare epidemiology of america (SHEA) research network and affiliated US-based hospitals[J]. Infection Control and Hospital Epidemiology, 2026: 1-8.
[37] Fan Z, Zhang Y, Wu Y, et al. Public understanding and misuse of antibiotics in China: insights from a comprehensive national survey[J]. The Lancet Regional Health. Western Pacific, 2025, 62: 101682.
[38] Sharma P, Sidhu B, Reddy A, et al. Artificial intelligence in anesthesia: comparison of the utility of ChatGPT v/s google gemini large language models in pre-anesthetic education: content, readability and sentiment analysis[J]. BMC Anesthesiology, 2025, 25: 574.
[39] Toohey K, Quince Z, Walker F, et al. Integrating generative AI in health education: a scoping review and implementation framework[J]. Medical Science Educator, 2025.
[40] Zhou M, Pan Y, Zhang Y, et al. Evaluating AI-generated patient education materials for spinal surgeries: comparative analysis of readability and DISCERN quality across ChatGPT and deepseek models[J]. International Journal of Medical Informatics, 2025, 198: 105871.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Additionalfile1.docx
- Additionalfile2.docx
- Additionalfile3.docx