Asia Pacific Science Press

Asia Pacific Journal of Clinical Medical Research
ISSN(O):3079-8345
ISSN(P):3079-8337
Vol. 1 No.4 (2025)

A Comparative Analysis of the Readability and Information Quality of the Chinese and English Versions of Educational Materials for Thoracic Surgery Patients Generated by DeepSeek, Grok-3 and ChatGPT

Shiyu Wang, Yuan Yu*

Cancer Hospital Thoracic Surgery, Cancer Hospital Chinese Academy of Medical Sciences and Peking Union Medical College, Panjiayuan South Lane, Chaoyang District, Beijing100021, China

Corresponding author: Yuan Yu, zlyyyuyuan@163.com

Abstract: Objective: To comparatively analyze the readability and information quality of the educational materials for patients undergoing thoracoscopic lobectomy in both Chinese and English versions generated by three mainstream Large Language Models (LLMS), namely DeepSeek, Grok-3 and ChatGPT, Provide evidence-based basis for the clinical selection of AI-assisted educational tools.**Method:** A cross-sectional study design was adopted, with "education for patients undergoing thoracoscopic lobectomy" as the core requirement. Standardized Chinese and English prompts were designed to drive each of the three models to generate 3 independent educational materials (a total of 18, 9 in Chinese and 9 in English). The readability was evaluated using the internationally recognized readability assessment tools (English: Flesch-Kincaid Grade Level, FKGL; Flesch Reading Ease, FRE; Chinese: average sentence length), and the DISCERN scale was used to evaluate the quality of information. The differences among the three models were compared by the Kruskal-Wallis H test, the differences between the Chinese and English versions were analyzed by the paired sample t-test, and the reliability of the raters was tested by the intraclass correlation coefficient (ICC).**Result:** 1. Readability: In the English version, DeepSeek V3 had the highest FRE score (80.36±1.18) and the lowest FKGL score (4.83±0.12), which was significantly better than ChatGPT-o3 (FRE: 67.36±0.74, FKGL:) 6.56±0.36) and Grok3 (FRE: 45.67±1.65, FKGL: 11.93±0.17) ($P<0.05$); Among the Chinese versions, Grok3 had the shortest average sentence length (17.74±1.02 characters), which was significantly better than ChatGPT-o3 (27.81±1.47 characters) and DeepSeek V3 (26.75±1.18 characters) ($P<0.05$).2. Information quality: The reliability of the raters was excellent (ICC=0.92, 95% CI: 0.925-0.998, $P<0.001$); The DISCERN total scores of the Chinese and English versions of the three models were all at the "good - excellent" level (59.00-71.17 points). Among them, the total scores of the Chinese and English versions of ChatGPT-o3 were the highest (English: 71.17±1.17, Chinese: 70.50±0.55), and Grok3 was the lowest (English: (63.17±0.94, Chinese: 59.00±0.89), and the difference between groups was statistically significant ($P<0.05$).**Conclusion:** Among the educational materials for thoracoscopic lobectomy generated by the three LLMS, the English version of DeepSeeking V3 has the best readability, the Chinese version of Grok3 has outstanding reading fluency, and the comprehensive performance of the Chinese and English versions of ChatGPT-o3 is balanced. The Chinese version still needs to be optimized in terms of terminology consistency and information details. When applying it in clinical practice, the model should be selected in combination with language requirements, and the content generated by AI should be professionally reviewed.

Published: Oct. 24, 2025
DOI: https://doi.org/10.62177/apjcmr.v1i4.731

1.Background

Lung cancer, a malignant tumor with both high incidence and mortality rates worldwide, has approximately 2.2 million new cases and 1.8 million deaths each year, according to data from the World Health Organization[1]. Thoracic surgery remains a key treatment option for early-stage and some mid-stage lung cancers[2]。 The degree to which patients understand the disease, surgical risks and key points of rehabilitation before and after the operation directly affects treatment compliance and prognosis.

Against the backdrop of the rapid iteration of artificial intelligence technology, AI tools have been deeply integrated into People's Daily lives. This application has also extended to the medical and health field. AI models represented by ChatGPT and DeepSeek, with their powerful natural language processing capabilities, are gradually becoming new tools for medical workers to assist in diagnosis and treatment and patient health management[3]。 AI tools, with their advantage of rapidly generating customized content, have provided a new path for the production of patient educational material. Research shows that many users consult large language models for medical advice, regardless of whether they have a formal clinical background[5]。 The systematic review[6] included 23 studies and found that 87% of them focused on the application of AI in surgical planning, while only 3 involved patient education. Recent reviews on AI in medical communication and patient education also generally pointed out that research in this field is still in its early stages. Especially, there is a gap in the generation and verification of specialized and personalized content[7].Although the application research of LLM in the medical and health field is increasing day by day, the research focusing on the generation of surgical education materials for specific specialties and comparing the performance of different mainstream models is still insufficient [8]. Research shows that when dealing with professional issues related to thoracic surgery, the GPT-4 version of ChatGPT demonstrates a high accuracy rate in self-education and self-assessment tests, reflecting its potential in understanding medical knowledge [9]. However, the current research has two limitations: First, there is a scarcity of studies focusing on thoracic surgery, especially thoracoscopic lobectomy. The core educational points such as the "minimally invasive characteristics" and "postoperative respiratory management" of this surgical procedure are significantly different from those in other surgical fields. Second, there is a lack of comparative analysis between the Chinese and English versions. With the increase in cross-border medical care and the medical needs of foreign patients, the demand for educational materials in both Chinese and English is becoming increasingly urgent. However, it is not yet clear whether there are differences in the generation quality of AI in different language environments. Therefore, this study aims to fill the above gap. The core purpose is to compare the readability differences between Chinese and English educational materials for thoracoscopic lobectomy generated by DeepSeek, Grok-3, and ChatGPT. 2) Evaluate the information quality (accuracy, completeness, clinical relevance, etc.) of the content generated by the three models; 3) Analyze the interactive influence of AI model types and language versions on the quality of educational materials.

2.Method

2.1 Research Design

A cross-sectional study design was adopted, and the research subjects were the educational materials for patients undergoing thoracoscopic lobectomy generated by three types of LLMS. To reduce the randomness of a single generation, each model generates three independent materials based on the same Prompt, ultimately forming an 18-sample library of "3 models ×2 languages ×3 materials".

2.2 AI Model Selection and Prompt Design

In this study, three cutting-edge large language models, namely DeepSeek V3 (DeepSeek AI), Grok-3 (xAI), and ChatGPT-o3 mini (OpenAI), were selected to generate educational materials for thoracic surgery. They are one of the most powerful artificial intelligence models in use worldwide. The Grok3 model from xAI, as a brand-new version released in 2025, focuses on enhancing efficient inference capabilities and is particularly suitable for portable devices and industrial edge computing

scenarios[10]. DeepSeek, as an emerging open-source model in China, has attracted widespread attention due to its efficient inference performance and excellent Chinese processing capabilities. The ChatGPT-o3 mini developed by OpenAI continues the leading position of this series of models in the field of natural language generation, achieving a balance between the efficiency and popularity of knowledge output in educational scenarios[11]. They can be used for free, enabling patients to easily access health information.

To ensure input consistency, the Chinese and English Prompt contents must strictly correspond. The core requirements include:

Target population: Patients undergoing thoracoscopic lobectomy.

Content scope: Surgical principles (minimally invasive advantages), preoperative preparations (smoking cessation, pulmonary function training, etc.), intraoperative procedures (anesthesia methods, operation duration), postoperative recovery (pain management, getting out of bed and moving around), complication prevention (atelectasis, bleeding, etc.), follow-up plans.

Language requirements: Easy to understand, avoid piling up professional terms (necessary terms should be accompanied by explanations).

English Prompt example "Generate patient education materials for video-assisted thoracoscopic lobectomy (VATS). The content must include: 1) VATS principle (minimally invasive advantages); 2) preoperative preparation (smoking cessation, pulmonary function training); 3) intraoperative process (anesthesia type, operation duration); 4) postoperative recovery (pain management, ambulation); 5) complication prevention (atelectasis, bleeding); 6) follow-up plan. The language should be easy to understand for patients with junior high school education or above, and professional terms (e.g., 'thoracoscope') must be explained simply."

2.3 Evaluation Tools

2.3.1 Readability Evaluation

English Reading materials, readability analysis was conducted using Flesch Reading Ease (FRE) and Flesch-Kincaid Grade Level (FKGL), among which the FRES score ranged from 0 to 100 (the higher the score, the easier it is to read). The SMOG index reflects the years of education required to understand the text (for example, an index of 10 indicates approximately the reading level of Grade 10). All indicators are calculated through the Readable online tool.

The average sentence length of the Chinese version of the missionary materials is compared.

2.3.2 Information Quality Assessment

DISCERN was developed by D. Charnock as an instrument to analyze the quality of health information[12]. The DISCERN tool was used for information quality assessment, which included three dimensions: reliability (8 items), treatment details (7 items), and overall quality (1 item). Each item was scored on a scale of 1 to 5 points (total score 16 to 80 points). The scoring criteria are defined as: >70 points (excellent), 60-69 points (good), 50-59 points (average), and <50 points (poor).

2.4 Data Collection and Statistical Analysis

From September 1st to September 15th, 2025, materials will be generated through the official API interfaces of each model. After extracting the text, it will be imported into the evaluation tool to calculate the readability index. Two reviewers trained by DISCERN (deputy chief nurses of thoracic surgery with more than 10 years of working experience and nursing education experts) independently scored and independently completed the information quality scoring. If the score difference was greater than 1 point, consensus was reached through discussion. SPSS 26.0 software was used, and the measurement data were expressed as "mean ± standard deviation (x±s)". The differences among the three models were analyzed using the Kruskal-Wallis H test (for non-normally distributed data), the differences between the Chinese and English versions were analyzed using the paired sample t-test, the reliability of the raters was analyzed using the intraclass correlation coefficient (ICC), and the correlation analysis was analyzed using the Pearson correlation coefficient. The test level $\alpha=0.05$.

3.Results

3.1 English Readability Analysis

There were significant differences in FRE and FKGL among the English versions of the three models (all $P<0.05$). The specific results are shown in Table 1: The FRE of DeepSeek V3 was significantly higher than that of ChatGPT-o3 and Grok3,

while FKGL was significantly lower than that of ChatGPT-o3 and Grok3. The FRE of ChatGPT-o3 was significantly higher than that of Grok3, and the FKGL was significantly lower than that of Grok3 (P=0.014). The readability ranking of the English version is indicated as: DeepSeek-V3>ChatGPT-o3>Grok3.

Table 1 Comparison of readability metrics of English educational materials generated by three AI models (x±s)

Model	Sample Size	Flesch Reading Ease (FRE)	Flesch-Kincaid Grade Level (FKGL)
Chatgpt-o3	3	67.36±0.74	6.56±0.36
Deepseek-V3	3	80.36±1.18	4.83±0.12
Grok3	3	45.67±1.65	11.93±0.17
H		7.82	
P		P<0.05	

3.2.2 Readability of Chinese version

There were significant differences in the average sentence lengths of the Chinese versions of the three models (H=7.20, P<0.05), and the specific results are shown in Table 2. The average sentence length of Grok3 is significantly shorter than that of ChatGPT-o3 and DeepSeek V3. The average sentence length of DeepSeek V3 was significantly shorter than that of ChatGPT-o3 (P=0.014). It indicates that the ranking of reading fluency of the Chinese version is: Grok3>DeepSeek V3>ChatGPT-o3

Table 2 Comparison of Readability Indicators of Educational Materials in Chinese Versions of Three AI Models (x±s)

Model	Sample Size	Average Sentence Length
Chatgpt-o3	3	27.81±1.47
Deepseek-V3	3	26.75±1.18
Grok3	3	17.74±1.02
H		7.2
P		P<0.05

3.3 DISCERN Information quality analysis

3.3.1 Rater reliability

Two reviewers scored the DISCERN scale of 18 materials with excellent reliability. The specific results are as follows:

DISCERN Total score ICC=0.92 (95% CI: 0.925, 0.998, P<0.001)

3.3.2 DISCERN Discern Total Score Comparison of Information quality

The DISCERN total scores of the Chinese and English versions of the three models were all at the “good - excellent” level, and there were significant differences among the groups (all P < 0.05). The specific results are shown in Table 3. In the English version, the total score of ChatGPT-o3 is significantly higher than that of Grok3, and that of DeepSeek V3 is significantly higher than that of Grok3. The differences between ChatGPT-o3 and DeepSeek V3 are nearly significant. In the Chinese version, the total score of ChatGPT-o3 is significantly higher than that of Grok3, and that of DeepSeek V3 is significantly higher than that of Grok3. The differences between ChatGPT-o3 and DeepSeek V3 are nearly significant.

The paired sample t-test showed that there were no significant differences in the DISCERN total scores of Chinese and English among the three models (ChatGPT-o3: t=1.28, P=0.27; DeepSeek-V3: t=-1.85, P=0.15;) Grok3: t=2.31, P=0.10), but the total score of the English version of Grok3 (63.17±0.94) was higher than that of the Chinese version (59.00±0.89), suggesting that the quality stability of its Chinese information was relatively weak.

Model	English Version	Chinese Version	English Inter-group
Chatgpt-o3	71.17±1.17	70.5±0.55	t=1.28, P=0.27
Deepseek-V3	66.00±0.89	67.83±0.75	t=-1.85, P=0.15
Grok3	63.17±0.94	59.00±0.89	t=2.31, P=0.10
H(English)	9.23		
P(English)	0.017		
H(Chinese)	9.87		
P(Chinese)	0.007		

4.Discussion

4.1 Differences in readability of Materials generated by the three LLMS and the reasons

This study found that the three LLMS demonstrated significant "language specificity" in terms of readability between the Chinese and English versions: DeepSeek V3 was the best in the English version, and Grok3 was the best in the Chinese version. This result is closely related to the characteristics of the training data of the models and the direction of language optimization.

The English version of DeepSeek V3 has outstanding readability, which may be attributed to its pre-training optimization on English medical texts. The training data of this model contains a large number of English patient education manuals, such as the public educational materials of Mayo Clinic and Johns Hopkins Hospital, and has undergone special fine-tuning for "popularization of medical information", which can precisely control the sentence length and vocabulary difficulty. The average sentence length of the Chinese version of Grok3 is the shortest, which is speculated to be related to its core positioning of "efficient reasoning". This model prioritizes the "short sentence splitting" strategy when generating Chinese, which, although it enhances fluency, may also lead to a slight decrease in content coherence.

The readability of the Chinese and English versions of ChatGPT-o3 is balanced, which is in line with its positioning as a "general-purpose LLM". The training data of this model covers multiple languages and fields. It performs stably in the balance of "readability - professionalism", but is slightly inferior to DeepSeek V3 (English) and Grok3 (Chinese) in the extreme optimization of a single language. The English version of Grok3 has the poorest readability, mainly because the content generated by this model contains unexplained professional terms and the sentence structure is complex, beyond the comprehension ability of patients.

4.2 Differences in Information Quality of Materials Generated by the Three LLMS and Clinical Implications

In terms of information quality, the overall performance of the Chinese and English versions of ChatGPT-o3 is the best, followed by DeepSeek V3, and Grok3 is the worst. This result is directly related to the model's medical knowledge reserve and content generation logic. The training data of ChatGPT-o3 contains a vast amount of medical literature and clinical diagnosis and treatment norms, and can accurately generate content that conforms to clinical consensus, such as quitting smoking two weeks before surgery and getting out of bed and moving around 24 hours after surgery. However, the quality of Grok3 information is insufficient mainly due to the relatively low proportion of medical and biomedical data in it, and there is a problem of "oversimplification", such as only describing the symptoms of atelectasis as "breathing difficulties", without mentioning key accompanying symptoms such as "chest pain and cough".

In addition, there are still inconsistent issues in the use of terms in the Chinese version. For instance, "thoracoscopy" is sometimes expressed as "chest wall endoscope", reflecting the insufficiency of LLM in standardizing Chinese medical terms. It is suggested that in subsequent studies, the "Uniform Requirements for Chinese Terminology" be added to the Prompt, and at the same time, a "Common Chinese Terminology Database for Thoracic Surgery" be established to provide a basis for terminology norms for AI-generated content.

4.3 Research Limitations and Future Directions

This study has certain limitations. First, the sample size is relatively small, with only three pieces of material generated for each model, which may lead to random errors. Subsequently, the sample size can be expanded to ten pieces of material for each model to enhance the extrapolation of the results. Second, the subjective evaluation of patients was not included. In the future, patients undergoing thoracoscopic lobectomy can be invited to rate the "understanding" and "practicality" of the materials, and a "subject-objective" combined assessment system can be formed by combining objective indicators. Thirdly, the impact of "version updates" on the model was not taken into account. LLMS have a fast iteration speed, and regular updates and research are needed in the future to track changes in model performance.

Future research can be carried out from two aspects: One is to explore the collaborative editing model of "AI + clinical doctors", where AI generates the initial draft and doctors supplement professional details to improve the quality of materials; Second, optimize the Prompt design for specific groups, such as elderly patients and foreign patients, to generate more personalized educational materials and meet the diverse clinical needs.

5.Conclusion

There are significant differences in readability and information quality among the educational materials for patients undergoing thoracoscopic lobectomy generated by the three LLMS: DeepSeeking V3 has the best readability in the English version; Grok3 has outstanding reading fluency in the Chinese version; and ChatGPT-o3 has a balanced overall performance in both Chinese and English versions. The consistency of terms in the Chinese version still needs to be prioritized for optimization. When applying in clinical practice, the model should be selected based on language requirements. For English scenarios, DeepSeek V3 is preferred; for Chinese scenarios, Grok3 can be chosen but the information quality needs to be reviewed. For dual-language scenarios, ChatGPT-o3 is preferred, and the AI-generated content should be reviewed and supplemented by thoracic surgery professionals. Conclusion 5: Ensure the Safety and effectiveness of materials

There are significant differences in readability and information quality among the educational materials for patients undergoing thoracoscopic lobectomy generated by three types of LLMS: In the English version, DeepSeek V3 has the best readability. In the Chinese version, Grok3 has outstanding reading fluency. The overall performance of ChatGPT-o3 in both Chinese and English versions is balanced (with a balance between readability and information quality). The consistency of terms in the Chinese version still needs to be prioritized for optimization. When applying in clinical practice, models should be selected based on language requirements (DeepSeek V3 is preferred for English scenarios, Grok3 for Chinese scenarios but the quality of information needs to be reviewed, and ChatGPT-o3 is preferred for multilingual scenarios). The AI-generated content should be reviewed and supplemented by thoracic surgery professionals to ensure the safety and effectiveness of the materials.

Funding

This work was supported by the National High Level Hospital Clinical Research Funding (80102022501).

Conflict of Interests

The author(s)declare(s) that there is no conflict of interest regarding the publication of this paper.

Reference

[1] Bray, F., Laversanne, M., Sung, H., Ferlay, J., Siegel, R. L., Soerjomataram, I., & Jemal, A. (2024). Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: A Cancer Journal for Clinicians, 74(3), 229-263. https://doi.org/10.3322/caac.21834

[2] Jiao, W., Zhao, L., Mei, J., Zhong, J., Yu, Y., Bi, N., ... & Gao, S. (2025). Clinical practice guidelines for perioperative multimodality treatment of non-small cell lung cancer. Chinese Medical Journal (English Edition). https://doi.org/10.1097/cm9.0000000000003635

[3] Lee, P., Bubeck, S., & Petro, J. (2023). Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. The New England Journal of Medicine, 388(13), 1233-1239. https://doi.org/10.1056/NEJMsr2214184

[4] Thirunavukarasu, A. J., Ting, D. S. J., Elangovan, K., Gutierrez, L., Tan, T. F., & Ting, D. S. W. (2023). Large language models in medicine. Nature Medicine, 29(8), 1930-1940. https://doi.org/10.1038/s41591-023-02448-8

[5] Choudhury, A., Shahsavar, Y., & Shamszare, H. (2025). User intent to use DeepSeek for healthcare purposes and their trust in the large language model: Multinational survey study. JMIR Human Factors. https://doi.org/10.2196/72867

[6] Bhushan, R., & Grover, V. (2024). The advent of artificial intelligence into cardiac surgery: A systematic review of our understanding. Brazilian Journal of Cardiovascular Surgery, 39(5), e20230308. https://doi.org/10.21470/1678-9741-2023-0308

[7] Denecke, K., May, R., & Rivera Romero, O. (2024). Potential of large language models in health care: Delphi study. Journal of Medical Internet Research, 26(5), e52399. https://doi.org/10.2196/52399

[8] Ayers, J. W., Poliak, A., Dredze, M., Leas, E. C., Zhu, Z., Kelley, J. B., Faix, D. J., Goodman, A. M., Longhurst, C. A., Hogarth, M., & Smith, D. M. (2023). Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. JAMA Internal Medicine, 183(6), 589-596. https://doi.org/10.1001/jamainternmed.2023.1838

[9] Khalpey, Z., Kumar, U., King, N., Abraham, A., & Khalpey, A. H. (2024). Large language models take on cardiothoracic surgery: A comparative analysis of the performance of four models on American Board of Thoracic Surgery exam questions in 2023. Cureus, 16(7), e65083. https://doi.org/10.7759/cureus.65083

[10] XAI. (2025). Grok3: Redefining AI Capabilities. Retrieved from https://xai.com/grok3

[11] OpenAI. (2024). ChatGPT Technical Report. Retrieved from https://openai.com/research/chatgpt

[12] Charnock, D., Shepperd, S., Needham, G., & Gann, R. (1999). DISCERN: An instrument for judging the quality of written consumer health information on treatment choices. Journal of Epidemiology & Community Health, 53(2), 105-111. https://doi.org/10.1136/jech.53.2.105