Postprint: Analysis of Information Quality and Readability of Thyroid Cancer Question Sets Based on Douyin Index in Large Language Models

Authors: Xue Mengyuan, Peng Yinghua, Ning Yanting, Ma Heng, Zhao Bohui, Huang Yingtong, Ning Yanting

Date: 2025-07-14T16:57:32+00:00

Abstract

Background: As an emerging technology, large language models are gradually gaining recognition and application among the general public. Thyroid cancer is a common type of malignant tumor in China, and patients have a high demand for thyroid cancer popular science information; however, there is still no domestic research analyzing the information quality and readability of response texts in the thyroid cancer domain within large language models. Objective: To evaluate and compare the information quality and readability of response texts to thyroid cancer-related questions generated by domestic large language models (LLMs). Methods: Based on the Douyin Index, 25 thyroid cancer questions were selected as a question set, and response texts were generated using DeepSeek (DeepSeek-R1-0120), Tongyi Qianwen (qwen-max-2025-01-25), and Zhipu Qingyan (GLM-4Plus), respectively. Cosine similarity was employed to calculate the similarity of texts generated at different time points to assess model stability. The modified health information quality evaluation tool (mDISCERN) was used to evaluate information quality, combined with a Chinese readability calculation formula to assess text readability. Differences in information quality of response texts among models were explored through clustering heatmaps, principal component analysis, Friedman test, and signed rank sum test, while Pearson correlation analysis was used to investigate the association between information quality and readability. Results: Text similarity evaluation results showed that for DeepSeek, moderately similar texts accounted for 12% and highly similar texts accounted for 88%; for Tongyi Qianwen and Zhipu Qingyan, highly similar texts accounted for 100% of the two response texts. Comparisons of information quality and readability among the three models revealed statistically significant differences ($P<0.001$). DeepSeek

was superior to other models in information quality (Z=35.396, P<0.001), but had relatively poor readability (R=7.525$±$1.006). Tongyi Qianwen and Zhipu Qingyan had similar information quality, but Zhipu Qingyan was better at responding to question set clusters 2 and 3, while Tongyi Qianwen was better at responding to question set cluster 1. Information quality was negatively correlated with readability (r=0.370, P=0.010). Conclusion: Domestic large language models can provide basic health popular science for thyroid cancer patients, but there are inaccuracies in generated content and artificial intelligence (AI) hallucinations. When patients actually use large language models (LLMs) to obtain health information, they should comprehensively consider response texts from different platforms and doctors' advice. In terms of models, it is necessary to balance information professionalism and popularization, and establish a medical content safety review mechanism to ensure information accuracy and professionalism.

Full Text

Analysis of Information Quality and Readability of Thyroid Cancer-Related Information in Large Language Models Based on Douyin Index

XUE Mengyuan[1], PENG Yinghua[1], NING Yanting[2]*, MA Heng[1], ZHAO Bohui[3], HUANG Yingtong[1]

[1]Department of Head and Neck Surgery, National Cancer Center/National Clinical Research Center for Cancer/Cancer Hospital & Shenzhen Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Shenzhen 518116, China
[2]Department of Nursing, National Cancer Center/National Clinical Research Center for Cancer/Cancer Hospital & Shenzhen Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Shenzhen 518116, China
[3]Department of Thoracic Surgery, National Cancer Center/National Clinical Research Center for Cancer/Cancer Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing 100021, China

Corresponding author: NING Yanting, Associate chief nurse; E-mail: ningyanting@cicams-sz.org.cn

---

Abstract

Background: Large language models (LLMs) are increasingly gaining public recognition and adoption as an emerging technology. Thyroid cancer is a prevalent malignancy in China, with patients demonstrating high demand for accessible scientific information about the disease. However, no studies have yet evaluated the information quality and readability of LLM-generated responses in the domain of thyroid cancer within the Chinese context.

Objective: To evaluate and compare the information quality and readability of responses generated by domestic large language models (LLMs) to thyroid cancer-related queries.

Methods: We identified 25 thyroid cancer questions based on Douyin Index data and generated response texts using DeepSeek (DeepSeek-R1-0120), Qwen (qwen-max-2025-01-25), and GLM (GLM-4Plus). Cosine similarity was employed to calculate the similarity of texts generated at different time points, thereby assessing model stability. The modified Health Information Quality Assessment Tool (mDISCERN) was used to evaluate information quality, while readability was assessed using a Chinese readability calculation formula. Cluster heatmaps, principal component analysis (PCA), Friedman tests, and signed rank tests were conducted to explore differences in response quality across models. Pearson correlation analysis was performed to examine the relationship between information quality and readability.

Results: Text similarity evaluation revealed that DeepSeek responses were moderately similar in 12% of cases and highly similar in 88%, while Qwen and GLM achieved 100% high similarity across their two response generations. Comparative analysis of information quality and readability across the three models showed statistically significant differences ($P<0.001$). DeepSeek demonstrated superior information quality ($Z=35.396$, $P<0.001$) but relatively poorer readability ($R=7.525\pm1.006$). Qwen and GLM exhibited comparable information quality, though GLM performed better on question clusters 2 and 3, while Qwen excelled on cluster 1. Information quality was negatively correlated with readability ($r=0.370$, $P=0.010$).

Conclusion: Domestic LLMs can provide basic health education for thyroid cancer patients, but suffer from content inaccuracies and AI hallucinations. Patients should comprehensively evaluate responses from different platforms alongside medical advice when using LLMs for health information. Model development requires balancing professional accuracy with accessibility while establishing robust medical content safety review mechanisms to ensure information accuracy and professionalism.

Keywords: Large language models; Thyroid cancer; Information quality; Readability analysis; Medical artificial intelligence

---

Introduction

According to 2022 epidemiological data from the National Cancer Center, new thyroid cancer cases in China reached 466,100, ranking third among all malignancies and showing an upward trend [1]. Patients experiencing treatment-related symptoms such as throat pain and limb numbness urgently require high-quality health information to inform their decision-making [2]. Large language models (LLMs), as frontier technologies in artificial intelligence, are gradually becoming

general-purpose intelligent tools in medical contexts through their capabilities in contextual learning, reasoning, and decision-making [3]. Their prompt-based question-answering abilities have demonstrated application potential in teaching, geriatric care, traditional Chinese medicine, and ophthalmic nursing [4-7]. One ophthalmology study found that LLM chatbot responses to patient eye health questions were comparable in quality to written recommendations from ophthalmologists [4]. However, domestic research evaluating the readability and information quality of LLM responses to thyroid cancer-related questions remains limited.

As of June 2024, China's short-video user base reached 1.056 billion (95.5% of all internet users) [8], with Douyin (the Chinese version of TikTok) averaging over 600 million daily active users. Information acquisition through short videos is growing rapidly among netizens. This study focuses on mainstream domestic LLMs (Qwen, GLM, and DeepSeek), constructing an evaluation question set based on "thyroid cancer" trending topics from Douyin's massive data platform. Using the modified Health Information Quality Assessment Tool (mDISCERN) [9] and readability calculation formulas, we conducted horizontal comparisons of response differences across the three LLMs to assess the safety, effectiveness, and feasibility of AI-assisted clinical applications, thereby reducing the risk of misleading decision-making.

Methods

1.1 Question Source Jùliàng Suànshù (Massive Engine's data analytics platform) provides content consumption trend analysis based on Douyin user behavior data. The Douyin Index measures keyword popularity, with relevance rankings based on the strength of association between related terms and the target keyword, displaying the top 10 most relevant associated terms. We collected thyroid cancer-related content from January to December 2023, obtaining 5,064 raw entries. After removing 4,028 duplicates and 749 entries with incomplete semantics or image-only content, we identified representative and research-valuable questions. Two researchers collaboratively curated the final question set, with a third researcher adjudicating any disagreements, resulting in a set of 25 questions (Table 1).

1.2 LLM Selection GLM-4Plus was developed by Tsinghua University [10], Qwen by Alibaba [11], and DeepSeek by the DeepSeek team [12]—all being popular LLMs among Chinese users. We used these models' ChatGLM GLM-4Plus, qwen-max-2025-01-25, and DeepSeek-R1-0120 versions to generate responses. New accounts were created for each model to ensure no prior history, and fresh conversations were initiated for each question to avoid context contamination. All responses were collected on February 9, 2025, and exported to Word for analysis.

1.3 Evaluation Metrics **1.3.1 Information Quality Assessment:** Information quality is defined as the degree to which information meets user needs [13]. The DISCERN instrument, widely used for evaluating health information reliability and quality across various medical domains [14], was adapted as mDISCERN. This version includes items 1-8 from the original DISCERN questionnaire for reliability assessment, scored on a 5-point scale (1=completely unmet to 5=completely met), with total scores ranging from 8-40. Scores of 8-15 indicate poor quality, 16-31 moderate, and 32-40 excellent quality, with validated applicability for LLM response evaluation [15]. Two clinical experts performed the assessment, blinded to model identity to ensure objectivity. Discrepancies were resolved by a third clinical expert through discussion and consensus.

1.3.2 Readability Calculation: Text readability refers to the reading comprehension level required to understand written materials, a critical factor in health information comprehension [16]. We employed the readability formula adapted by Qin et al. [17] for medical information assessment:

$$R = 17.5255 + 0.0024 \times \text{total words} + 0.04415 \times \text{average sentence length} - 18.3344 \times (1 - \text{proportion of medical term}$$

The R-value represents health information readability level, where lower values indicate better readability. Total word count excludes punctuation. Average sentence length equals total words divided by complete sentences. Medical terminology proportion equals medical term characters divided by total characters [17].

Implementation: Total words and sentences were counted in Word documents. For medical terminology proportion, we first used the Language Technology Platform (LTP) for Chinese word segmentation [18]. We then used the LetPub Medical English Dictionary (2025 edition) as our terminology corpus, which includes 16,229 specialized terms across 46 medical disciplines. Excel's VLOOKUP function matched segmented words against this dictionary to count medical term characters, calculating the final proportion.

1.4 Stability Testing On June 21, 2025, we re-input the question set into the three LLMs. Cosine similarity measured the similarity between the two response sets. After removing punctuation, stop words, and special characters, texts were segmented and vectorized based on unique word lists. Cosine similarity (Rs) calculates the cosine of the angle between two vectors in vector space [19]; Rs approaching 1 indicates high similarity, while Rs near 0 indicates low similarity. Similarity was categorized as: very different (Rs<0.3), slightly similar (0.3≤Rs<0.5), moderately similar (0.5≤Rs<0.7), highly similar (0.7≤Rs<0.9), and nearly identical (0.9≤Rs$ $1), assessing whether LLMs provide consistent responses to identical questions over time.

1.5 Statistical Analysis Data analysis was performed using R version 4.4.3. Normally distributed continuous data were expressed as mean±standard deviation ($\bar{x}\pm s$) and compared using one-way ANOVA. Non-normally distributed data were presented as median (P25, P75) and analyzed using Friedman tests with post-hoc signed rank tests. Readability scores were calculated in Excel and visualized with bar charts. The "pheatmap" package in R was used to generate cluster heatmaps of information quality scores (with "scale=row" for row normalization, "clustering_{method}=complete", and Euclidean distance). Principal component analysis used the "prcomp" function, visualized with "ggplot2". Pearson correlation analysis examined the relationship between information quality and readability using "cor.test". Statistical significance was set at $P<0.05$.

Results

2.1 Search Distribution and TGI Index Analysis Douyin Index data for "thyroid cancer" from January-December 2023 showed regional concentration in search distribution: Guangdong (9.58%), Jiangsu (9.25%), Henan (7.42%), Shandong (6.86%), Zhejiang (6.70%), and Anhui (5.93%) accounted for 45.74% of total searches, while Tibet (0.20%) and Hong Kong (0.01) contributed less than 0.5%. Target Group Index (TGI) analysis revealed above-average attention (TGI>100) in Anhui (145), Beijing (140), Jiangsu (130), Shanghai (125), and Hubei (122). Provinces with below-average attention (TGI<100) included Guangdong (85), Sichuan (85), and Guangxi (71). Notably, Guangdong ranked first in distribution proportion (9.58%) but had a TGI of only 85, while Anhui, despite comprising 5.93% of searches, achieved the highest national TGI (145) (Figure 1 [Figure 1: see original paper]).

2.2 LLM Response and Stability Results Qwen exhibited factual errors in epidemiological responses, stating "thyroid cancer is a relatively rare cancer type, but it can indeed cause death"—contradicting current reports that thyroid cancer reached 466,100 new cases in 2022 with an age-standardized incidence rate of 24.64/100,000, ranking third among all malignancies after lung and colorectal cancers [1]. For questions regarding "postoperative dietary restrictions" and "postoperative diet recipes," Qwen and GLM only vaguely mentioned "avoiding irritating foods: spicy, fried, and heavily flavored foods may increase gastrointestinal burden," without specific guidance on fat intake that could risk chyle leakage.

Comparative analysis of the two response generations revealed decreasing similarity across all models. DeepSeek showed 12% moderate similarity and 88% high similarity, while Qwen and GLM achieved 100% high similarity (Table 2).

2.3 Information Quality Analysis mDISCERN evaluation followed by Friedman nonparametric tests revealed statistically significant differences in information quality and readability across the three models ($P<0.001$). DeepSeek

outperformed both GLM and Qwen in information quality (P<0.001), while GLM and Qwen showed comparable quality (P>0.05) (Table 3).

Cluster heatmap and PCA analysis of the 25 questions identified three question clusters and two model categories: DeepSeek as one category, and Qwen and GLM as another. DeepSeek achieved high quality scores (orange) across all 25 questions. GLM scored lower on questions in the upper portion (questions 20, 12, 6, 19, 1, 24, 5, 9), while Qwen scored lower on lower-portion questions (questions 21, 13, 25, 23, 10, 3, 18, 17, 4, 7, 22, 15, 2, 14, 8, 11). Notably, question 16 (long-term side effects of postoperative levothyroxine) showed divergent scores: moderate in DeepSeek, high in GLM, and low in Qwen, indicating substantial variation in responses to medication side effect queries (Figure 2 [Figure 2: see original paper]A).

Based on heatmap clustering, questions were grouped into: Cluster 1 (questions 20, 12, 6, 19, 1, 24, 5, 9), Cluster 2 (questions 21, 13, 25, 23, 10, 3, 18), and Cluster 3 (questions 17, 4, 7, 22, 15, 2, 14, 8, 11). PCA revealed that Qwen and GLM were similar on PC1, while DeepSeek differed significantly. All three models showed substantial differences on PC2. PCA of the question set showed no major differences on PC1 across the three clusters but clear separation on PC2, consistent with heatmap clustering results (Figure 2 [Figure 2: see original paper]B, C).

2.4 Readability Analysis Readability calculations showed DeepSeek scores were generally higher (worse readability) than Qwen and GLM (Figure 3 [Figure 3: see original paper]). The difference was statistically significant (F=2.533, P<0.001) with a moderate effect size (Cohen's f=0.59) (Table 3).

Descriptive analysis revealed distinct text characteristics: DeepSeek generated the most information-dense content ($885.92 \pm 220.49 characters$) $throughmulti-paragraphstructures$ ($20.28 \pm 6.03 sentences$) $andlongsentences$ ($45.42 \pm 10.46 characters/sentence$), $containingt$ (Table 4).

2.5 Quality-Readability Correlation Analysis Pearson correlation analysis revealed a positive correlation between information quality and readability scores (r=0.370, P=0.010), indicating that higher information quality was associated with higher readability scores (worse readability).

Discussion

LLMs have demonstrated effectiveness in improving healthcare task performance and satisfaction among older adults [5]. Compared to international models, China's LLMs developed later, though evaluations of domestic models (Wenxin Yiyan, iFlytek Spark) show comparable accuracy to ChatGPT in nursing fundamentals [6]. However, research on specific diseases (e.g., erectile dysfunction) reveals that LLMs (ChatGPT, Bard, Bing, Ernie, Copilot) generally fail to meet readability standards, exhibit inconsistent information

quality, and contain quality issues [20]. Responding to the "14th Five-Year Plan" for national health development, this study focused on thyroid cancer to evaluate DeepSeek, Qwen, and GLM.

DeepSeek achieved optimal information quality but poorest readability, while Qwen and GLM showed higher readability but lower information quality. All models demonstrated decreased response similarity over time, with significant variation in responses to question 16 (levothyroxine side effects). This suggests patients should cross-reference multiple models and consult physicians to avoid decision-making errors based on single sources.

AI hallucination—generating plausible but factually incorrect content—remains an unavoidable challenge. Additionally, model knowledge cutoffs may exclude the latest clinical data, contributing to inaccurate responses. These safety concerns can impact patient behavior and medical decisions, potentially eroding trust and exacerbating doctor-patient conflicts. However, all three models included safety reminders to follow medical advice and seek timely treatment, somewhat mitigating risks.

This study has limitations. The sample size and disease coverage were limited; future research should expand question sets and disease types. Cosine similarity measures surface-level text similarity but doesn't assess consistency of key data (e.g., incidence rates, survival statistics) or completeness of risk warnings. Future studies should develop multi-dimensional evaluation tools covering safety, accuracy, readability, and stability for medical domains. Additionally, exploring model applications across different scenarios (education, triage, documentation, quality control) would be valuable.

In conclusion, this comprehensive evaluation of three LLMs for thyroid cancer health queries indicates that while all models can provide reproducible answers with good stability for medical education, each has distinct strengths and weaknesses. Safety risks and AI hallucinations remain non-negligible and may mislead patient health decisions. Patients must synthesize responses from multiple models with physician guidance. Therefore, LLM deployment for medical education requires balancing professionalism, accessibility, and stability while establishing effective safety review mechanisms to ensure compliance with medical ethics and clinical standards.

Author Contributions: XUE Mengyuan conceptualized and designed the study, performed statistical analysis, created figures and tables, drafted the manuscript, and takes overall responsibility. PENG Yinghua implemented the study. NING Yanting proposed the main research objectives, supervised quality control, and provided oversight. PENG Yinghua, MA Heng, and ZHAO Bohui designed the question set and evaluated information quality. PENG Yinghua and HUANG Yingtong revised the manuscript.

Conflict of Interest: The authors declare no conflicts of interest.

Funding: This study was supported by the Nursing Research Special Project of Shenzhen Hospital, Cancer Hospital, Chinese Academy of Medical Sciences (E010422009).

Data Availability: The datasets generated and analyzed during this study are available from the corresponding author upon reasonable request.

Ethical Approval: Not applicable.

Informed Consent: Not applicable.

Citation: XUE MY, PENG YH, NING YT, et al. Analysis of information quality and readability of thyroid cancer-related information in large language models based on Douyin index [J]. Chinese General Practice, 2025. DOI: 10.12114/j.issn.1007-9572.2025.0142. [Epub ahead of print]

References

[1] HAN BF, ZHENG RS, ZENG HM, et al. Cancer incidence and mortality in China, 2022[J]. J Natl Cancer Cent, 2024, 4(1): 47-53. DOI: 10.1016/j.jncc.2024.01.006.

[2] ZHANG J, LI LM, LI JY, et al. Survey on health information acquisition needs of thyroid cancer patients[J]. Nursing Research, 2019, 33(22): 3872-3878. DOI: CNKI:SUN:SXHZ.0.2019-22-011.

[3] ZHAO WX, ZHOU K, LI J, et al. A survey of large language models[J/OL]. arXiv, 2025[2025-04-01]. http://arxiv.org/abs/2303.18223. DOI: 10.48550/arXiv.2303.18223.

[4] BERNSTEIN IA, ZHANG YV, GOVIL D, et al. Comparison of ophthalmologist and large language model chatbot responses to online patient eye care questions[J]. JAMA Netw Open, 2023, 6(8): e2330320. DOI: 10.1001/jamanetworkopen.2023.30320.

[5] KHAMAJ A. AI-enhanced chatbot for improving healthcare usability and accessibility for older adults[J]. Alex Eng J, 2025, 116: 202-213. DOI: 10.1016/j.aej.2024.12.090.

[6] XU WB, ZHOU XP. Exploration of ChatGPT-like large language models in nursing curriculum assessment: Testing based on ChatGPT, Wenxin Yiyan, and iFlytek Spark[J]. China Medical Education Technology, 2024, 38(5): 567-571. DOI: 10.13566/j.cnki.cmet.cn61-1317/g4.202405005.

[7] XIANG BZ, WANG ZZ, ZHAO YS, et al. Performance and analysis of large language models in the Chinese medical practitioner qualification examination[J]. Traditional Chinese Medicine Education, 2025, 44(1): 1-6.

[8] The 54th Statistical Report on China's Internet Development[J]. Media Forum, 2024, 7(17): 121.

[9] MACLEOD MG, HOPPE DJ, SIMUNOVIC N, et al. YouTube as an information source for femoroacetabular impingement: A systematic review of video content[J/OL]. Arthroscopy: The Journal of Arthroscopic & Related Surgery, 2015, 31(1): 136-142. DOI: 10.1016/j.arthro.2014.06.009.

[10] GLM T, ZENG A, XU B, et al. ChatGLM: A family of large language models from GLM-130B to GLM-4 all tools[J/OL]. arXiv, 2024[2025-04-01]. http://arxiv.org/abs/2406.12793. DOI: 10.48550/arXiv.2406.12793.

[11] AZIZI OTHMAN. Technical report on qwen-2.5 max[J/OL]. 2025[2025-04-01]. https://rgdoi.net/10.13140/RG.2.2.31239.51363. DOI: 10.13140/RG.2.2.31239.51363.

[12] DEEPSEEK-AI, GUO D, YANG D, et al. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning[J/OL]. arXiv, 2025[2025-04-01]. http://arxiv.org/abs/2501.12948. DOI: 10.48550/arXiv.2501.12948.

[13] WANG RY, STRONG DM. Beyond accuracy: What data quality means to data consumers[J]. J Manag Inf Syst, 1996, 12(4): 5-33. DOI: 10.1080/07421222.1996.11518099.

[14] CHARNOCK D, SHEPPERD S, NEEDHAM G, et al. DISCERN: An instrument for judging the quality of written consumer health information on treatment choices[J]. J Epidemiol Community Health, 1999, 53(2): 105-111. DOI: 10.1136/jech.53.2.105.

[15] ONDER CE, KOC G, GOKBULUT P, et al. Evaluation of the reliability and readability of ChatGPT-4 responses regarding hypothyroidism during pregnancy[J]. Sci Rep, 2024, 14(1): 243. DOI: 10.1038/s41598-023-50884-w.

[16] ALBRIGHT J, DE GUZMAN C, ACEBO P, et al. Readability of patient education materials: Implications for clinical practice[J]. Appl Nurs Res, 1996, 9(3): 139-143. DOI: 10.1016/s0897-1897(96)80254-0.

[17] QIN Q, KE Q, DING SY. Readability calculation and empirical application of Chinese online health education information: A case study of food safety[J]. Modern Intelligence, 2020, 40(5): 111-121. DOI: 10.3969/j.issn.1008-0821.2020.05.014.

[18] CHE W, FENG Y, QIN L, et al. N-LTP: An open-source neural language technology platform for Chinese[J/OL]. arXiv, 2021[2025-04-01]. http://arxiv.org/abs/2009.11616. DOI: 10.48550/arXiv.2009.11616.

[19] LÜ YF. Research and application of text similarity algorithms for scientific project duplication checking[D/OL]. Institute of Disaster Prevention, 2024. DOI: 10.27899/d.cnki.gfzkj.2024.000048.

[20] ŞAHIN MF, ATEŞ H, KELEŞ A, et al. Responses of five different artificial intelligence chatbots to the top searched queries about erectile dysfunction: A comparative analysis[J]. J Med Syst, 2024, 48(1): 38. DOI: 10.1007/s10916-024-02056-0.

[21] XU CJ. Internet search engine data for tumor surveillance, early warning, and management[D]. Tianjin: Tianjin Medical University, 2021. DOI: 10.27366/d.cnki.gtyku.2021.000036.

[22] TAN XW, CHEN WF, WANG NN, et al. Query responses and effectiveness evaluation of domestic large language models on perioperative nursing and health education for prostate cancer[J/OL]. Chinese Journal of Andrology, 2024, 30(2): 151-156. DOI: 10.13263/j.cnki.nja.2024.02.010.

[23] ZHOU C, LIU P, XU P, et al. LIMA: Less is more for alignment[J/OL]. arXiv, 2023[2025-04-01]. http://arxiv.org/abs/2305.11206. DOI: 10.48550/arXiv.2305.11206.

[24] FU YT, QIU JP, ZHANG TY, et al. Quantitative analysis of information quality and readability in online health platforms[J]. Modern Intelligence, 2024, 44(3): 140-151. DOI: 10.3969/j.issn.1008-0821.2024.03.013.

[25] SWELLER J. Cognitive load during problem solving: Effects on learning[J]. Cogn Sci, 1988, 12(2): 257-285. DOI: 10.1016/0364-0213(88)90023-7.

[26] CHANDLER P, SWELLER J. Cognitive load theory and the format of instruction[J]. Cogn Instr, 1991, 8(4): 293-332. DOI: 10.1207/s1532690xci0804_2.

[27] PARK I, HER N, CHOE JH, et al. Management of chyle leakage after thyroidectomy, cervical lymph node dissection, in patients with thyroid cancer[J]. Head Neck, 2018, 40(1): 7-15. DOI: 10.1002/hed.24852.

[28] ZHOU M, PAN Y, ZHANG YY, et al. Evaluating AI-generated patient education materials for spinal surgeries: Comparative analysis of readability and DISCERN quality across ChatGPT and deepseek models[J]. Int J Med Inform, 2025, 198: 105871. DOI: 10.1016/j.ijmedinf.2025.105871.

[29] XIA SJ, HUA Q, MEI ZH, et al. Clinical application potential of large language model: A study based on thyroid nodules[J]. Endocrine, 2025, 87(1): 206-213. DOI: 10.1007/s12020-024-03913-1.

Tables and Figures