Contemporary Education Frontiers, 2026, Volume 4, Issue 5
https://bbwpublisher.com/journal/CEF
Online ISSN: 3029-1879
Print ISSN: 3029-1860

A Comparative Study on the Multidimensional Performance of Large Language Models in Long-text Medical Education in Hematology

Jie Wang[1], Zizhu Lian[2], Yulin Jiang[1], Zhiyi Li[1], Pengcheng He[1]*

[1]Department of Hematology, The First Affiliated Hospital of Xi'an Jiaotong University, Xi'an 710061, Shaanxi, China
[2]Department of Cardiovascular Surgery, The First Affiliated Hospital of Xi'an Jiaotong University, Xi'an 710061, Shaanxi, China

Abstract: *Background:* As large language models (LLMs) surpass the million-token context window, their potential in medical long-text education has become increasingly evident. However, systematic evaluation of performance differences among various models in hematology long-text teaching scenarios remains lacking. It is an urgent need to introduce large language models to revolutionize education. *Objective:* To compare the comprehensive performance of large language models in long-text teaching tasks in Hematology. *Methods:* Four categories of long-text teaching tasks (n = 24) were constructed, including instructional guidelines, progressive analysis of long cases, multiple rounds of Socratic questioning, and cross-document knowledge integration. Representative LLMs (GPT-5.5 and Kimi K2.6 as examples) were selected to complete these tasks. Nine hematology teaching invited for blind scoring to evaluate six dimensions: medical accuracy, teaching logic, long-text information integration ability, language adaptability, educational inspiration, and response completeness. Each dimension was scored from 0 to 20, with a total score ranging from 0 to 120. Long-text specific metrics, including information omission rate and cross-document consistency error were also quantified. *Results:* No significant difference was found in overall scores between the two models (100.01 ± 1.45 vs. 100.30 ± 1.52, $P = 0.623$), demonstrating a pattern of "convergent total scores with divergent dimensional performance." Dimensional analysis showed that GPT-5.5 outperformed Kimi K2.6 in medical accuracy (17.11 ± 0.32 vs. 16.89 ± 0.28, $P = 0.089$), linguistic appropriateness (16.71 ± 0.29 vs. 16.15 ± 0.41, $P = 0.018$), and educational heuristics (17.25 ± 0.31 vs. 15.58 ± 0.22, $P < 0.001$), whereas Kimi K2.6 performed better in pedagogical logic (17.05 ± 0.30 vs. 16.47 ± 0.28, $P = 0.042$), long-text information integration (17.42 ± 0.33 vs. 15.83 ± 0.35, $P < 0.001$), and response completeness (17.21 ± 0.35 vs. 16.64 ± 0.38, $P = 0.003$). Kimi K2.6 demonstrated a significantly lower information omission rate (4.2 ± 1.3% vs. 8.5 ± 2.1%, $P < 0.001$), and GPT-5.5 exhibited a notable "Lost in the Middle" phenomenon. *Conclusion:* In the long-text teaching in the Hematology Department, models with strong long-text integration capabilities are suitable for systematic lesson preparation and long-term case teaching, which can effectively reduce the rate of information omission and ensure the completeness of teaching content and the coherence of the time axis. Models with high educational heuristics are more suitable for clinical thinking training and multi-round interactive teaching. It is recommended to establish a three-layer application framework of "education demand-driven, model capability-matched, and teacher quality control as the bottom line". This will promote AI-assisted teaching to move

from “available” to “trustworthy” and “optimal use”.

Keywords: LLMs; long-text processing; medical education; Hematology Department; AI-assisted teaching

Online publication: May 26, 2026

1. Introduction

With the rapid advancement of artificial intelligence, large language models (LLMs) have demonstrated increasingly prominent value in the medical domain, encompassing medical education, information acquisition, clinical decision support, health science popularization, and related areas [1-4]. Compared with conventional search engines, LLMs are capable of delivering immediate and personalized responses to complex inquiries by virtue of their robust contextual understanding and natural language generation capabilities. Early research primarily focused on short-text medical question-answering tasks, evaluating the models’ medical accuracy and empathetic expressive ability [5, 6]. However, as the model context window has expanded from the early 4K tokens to million-token scales, the application scenarios of LLMs are undergoing a profound shift from “fragmented question answering” to “systematic teaching assistance” [7, 8].

The concept of long-text teaching originates from a critical reflection on traditional fragmented instructional models. As early as the beginning of the 21st century, with the exponential growth of medical knowledge and the marked increase in the volume and complexity of clinical guidelines, diagnostic and therapeutic norms, and evidence-based medicine literature, the medical education community began to focus on how to conduct systematic teaching based on complete original documents. After 2010, problem-based learning (PBL) and case-based learning (CBL) were gradually introduced to incorporate comprehensive analyses of long-duration disease cases [9-11]; nevertheless, constrained by the limited working memory capacity of the human brain and classroom time restrictions, it remains difficult for instructors to present complete original materials exceeding 10,000 words in a single session. Consequently, they often have to resort to highly condensed summaries or excerpts, which prevent students from accessing the complete logic and detailed evidence of the original literature. In recent years, with the widespread adoption of electronic health record systems and the opening of medical databases, the demand for long-text teaching in specialized fields such as hematology has become increasingly urgent; however, traditional teaching modalities encounter significant bottlenecks when processing ultra-long texts.

In the current context of medical education, the field faces multiple challenges [12], among which long-text instruction exhibits the following core shortcomings. First, the problem of information truncation is prominent. Instructors must manually screen and extract key content during lesson preparation, and subjective selection biases lead to information loss. In the field of hematology, in particular, critical details embedded in long texts—such as diagnostic thresholds for specific genotype subtypes in WHO classification standards or updates to NCCN guidelines—are prone to being overlooked. Second, cross-document integration efficiency is low. When teaching requires simultaneous reference to multiple versions of guidelines, textbooks, and recent literature, instructors spend considerable time on manual comparisons and struggle to precisely identify version discrepancies and contradictory statements across different sources. Third, the capacity for personalized adaptation is insufficient. Traditional long-text instruction adopts a “one-size-fits-all” model, making it difficult to dynamically adjust content depth and presentation format according to students’

cognitive levels. Fourth, interactive feedback is delayed. Questions that arise after students read long texts often cannot be addressed immediately and in depth during class, thereby impairing the internalization of knowledge.

The introduction of large language models into long-text instruction, as a novel teaching paradigm, can assist in systematic instructional design, progressive case analysis, multi-turn dialogic guidance, and cross-document knowledge integration, and holds substantial necessity and educational value [1, 13, 14]. First, large language models can function as "intelligent information fidelity preservers": with million-token context windows, they fully retain the entire content of original sources, preventing information truncation caused by manual extraction and ensuring a high degree of consistency between teaching materials and primary documents. Second, these models can act as "cross-source knowledge integrators," automatically identifying differences, updates, and contradictory statements across multiple documents and presenting them in a structured format, thereby greatly improving preparation efficiency and knowledge integration accuracy. Third, as "personalized cognitive scaffolds," large language models can generate targeted explanations in response to student queries, transforming abstract concepts in long texts into concrete examples adapted to different learning stages, thus facilitating a pedagogical shift from "textbook-centered" to "learner-centered" instruction. Finally, in multi-turn dialogues, the models can assume the role of "Socratic guides," stimulating students' critical thinking and clinical reasoning abilities through sequential questioning and counterfactual scenario construction.

Without the support of large language models, the current teaching model exhibits insurmountable structural deficiencies in long-text processing. Taking hematology as an example, under the traditional model, preparing a 45-minute lecture based on the complete CSCO guidelines typically requires 4–6 hours of manual reading, extraction, and reorganization, and full coverage of remote sections of the guidelines (e.g., recommendations for rare subtype diagnosis and treatment) remains difficult to guarantee. In long-case teaching, when students are presented with a complete clinical record exceeding 10,000 words, they are often provided with only a few preset "standard questions," precluding in-depth exploration of detailed changes in the disease course. In cross-version literature comparison teaching, manual comparison by instructors is not only time-consuming and labor-intensive but also prone to misjudgments due to fatigue. These deficiencies make it difficult to meet the deeper requirements of cultivating students' systematic knowledge construction and critical thinking abilities in hematology education. Therefore, exploring the application of large language models in long-text instruction is not a simple technological substitution but rather a pedagogical paradigm innovation addressing the deep-seated needs of medical education, and it holds significant practical value for improving the quality of medical talent cultivation [14].

We selected two representative contemporary large language models—GPT-5.5 (OpenAI, 2026) and Kimi K2.6 (Moonshot AI, 2026)—as case examples for comparative analysis. The former demonstrates outstanding performance in multimodal understanding, chain-of-reasoning generation, educational heuristic potential, and humanistic depth [15, 16]; the latter supports an ultra-long context of 2 million characters, offering theoretical advantages in complete document information extraction and cross-document integration [17]. However, empirical data are still lacking regarding the actual performance differences between the two in hematology long-text teaching scenarios, particularly with respect to pedagogical dimensions such as instructional logic, long-text information integration precision, and educational heuristic value. This study focuses on representative long-text teaching tasks in hematology and systematically compares the

performance differences between these two representative models, aiming to provide evidence-based guidance for the selection of digital teaching resources.

2. Methods

2.1. Study Design

This study employed a mixed-method approach combining quantitative scoring with qualitative analysis to compare comprehensive performance in long-text teaching tasks within hematology. The research encompassed overall and dimension-specific scoring, differential analysis across task types, and quantitative evaluation of long-text-specific indicators.

This study was approved by the Institutional Review Board of the hospital (No. XJTU1AF2026LSYY-0466) and was registered with the Medical Research Registration Information System (Registration No. MR-61-26-058628).

2.2. Long-Text Teaching Task Design

Based on the cognitive load characteristics and information processing complexity inherent in long-text teaching scenarios, four task types were established (designated T1–T4, where T stands for Task Type, intended to distinguish different teaching functions rather than represent a simple sequence of questions): T1 focused on systematic instructional design grounded in authoritative literature; T2 emphasized progressive deconstruction of longitudinal disease courses; T3 concentrated on dialogic cognitive guidance; and T4 centered on critical synthesis of multi-source information. A total of 24 specific tasks were constructed, detailed as follows:

T1 – Guideline-based instructional design (6 tasks): Complete guidelines or textbook chapters (8,000–15,000 words) were input, requiring the output of a 45-minute teaching plan covering the following diseases: acute myeloid leukemia (AML), diffuse large B-cell lymphoma (DLBCL), myelodysplastic syndromes (MDS), primary immune thrombocytopenia (ITP), complications following hematopoietic stem cell transplantation, and multiple myeloma (MM).

T2 – Progressive analysis of long-duration case histories (6 tasks): Long-term medical records (5,000–10,000 words) were input, with the requirement to design a stage-by-stage teaching guidance outline covering: Philadelphia chromosome-positive acute lymphoblastic leukemia (Ph+ ALL), acute promyelocytic leukemia (APL) complicated by disseminated intravascular coagulation (DIC), severe aplastic anemia (SAA), Waldenström macroglobulinemia (WM), aggressive natural killer (NK) cell leukemia, and hemophilia A.

T3 – Multi-round Socratic questioning (6 tasks): Three consecutive rounds of follow-up questions were administered to assess contextual consistency and error-correction capability, covering: iron deficiency anemia (IDA), chronic myeloid leukemia (CML), disseminated intravascular coagulation (DIC), autoimmune hemolytic anemia (AIHA), agranulocytosis with fever, and transfusion-related acute lung injury (TRALI).

T4 – Cross-document knowledge integration (6 tasks): Two to three documents from different sources (totaling 10,000–20,000 words) were simultaneously input, requiring the identification of discrepancies and their synthesis into teaching points, covering: AML risk stratification, MDS diagnostic criteria, second-line therapy for ITP, CML treatment response monitoring, positron emission tomography-computed tomography (PET-CT) response assessment in lymphoma, and gene therapy for hemophilia A.

2.3. Model Response Generation

Two representative contemporary large language models, GPT-5.5 (OpenAI, 2026) and Kimi K2.6 (Moonshot AI, 2026), were selected to complete all 24 tasks. Inputs were kept consistent, with no contextual memory or additional prompt engineering introduced. For T1, T2, and T4, a unified document-format input was adopted; for T3, inputs were entered sequentially round by round. Responses from both models were generated in Chinese, anonymized, and stripped of model-specific formatting markers before conversion to plain text.

2.4. Expert Scoring System

An evaluation team was composed of nine hematology specialists (three chief physicians, three teaching supervisors, and three senior attending physicians). The scoring framework encompassed six dimensions: medical accuracy, pedagogical logic, long-text information integration capability, linguistic adaptability, educational heuristic value, and response completeness. Each dimension was scored on a 0–20 scale, yielding a total score ranging from 0 to 120. Scoring was performed in a blinded manner. Prior to the formal experiment, a pilot study was conducted using T1-1 and T4-1 to calculate inter-rater consistency among the nine experts. The intraclass correlation coefficient [ICC(2,1)] across the six dimensions ranged from 0.78 to 0.91 (all > 0.75), indicating good inter-rater agreement and satisfying the requirements for subsequent statistical analysis.

The evaluation criteria for the six dimensions were established based on the following rationales: The medical accuracy dimension was anchored to the core requirement of "knowledge correctness" in medical education evaluation, ensuring that model outputs adhered to evidence-based medicine principles and the latest guideline consensus. The teaching logic dimension drew upon the "analysis" and "synthesis" levels of Bloom's taxonomy of educational objectives, assessing whether the organizational structure of the output followed cognitive progression patterns. The long-text information integration capability dimension was specifically designed to address the unique demands of long-text teaching, focusing on the model's capacity to preserve remote information from input documents and its ability to establish cross-paragraph associations. The language adaptability dimension referenced medical communication theory, evaluating performance in terms of medical terminology density, conciseness of expression, and audience appropriateness. The educational inspiration dimension was grounded in constructivist learning theory, assessing the model's ability to stimulate deep student thinking through questioning, analogies, and counterfactual scenario construction. The response completeness dimension, based on the systemic requirements of instructional design, evaluated whether the output covered all required knowledge points and teaching components specified by the task. This six-dimensional scoring system demonstrated good consistency (ICC > 0.75) in the pilot validation, effectively capturing the comprehensive performance of large language models in long-text teaching contexts.

2.5. Indicator Collection and Evaluation Process

Quantitative indicators included dimension-specific scores, total scores, response time, generated text length, and the word count-to-score ratio. Long-text specific indicators comprised: information omission rate (number of omitted key knowledge points / total number of key knowledge points), cross-document consistency errors (number of erroneous integration instances in T4 tasks), and contextual hallucination rate (number of fabricated contents not mentioned in the input). Qualitative analysis was conducted by two researchers who

coded teaching strategies and long-text error patterns, with disagreements adjudicated by a third party.

2.6. Statistical Analysis

Data organization and analysis were performed using GraphPad Prism (version 10.0) and R language (version 4.3.2). Normally distributed continuous data were expressed as mean ± standard deviation ($\bar{x} \pm s$), with between-group comparisons conducted using paired-sample t-tests; non-normally distributed data were analyzed using the Wilcoxon signed-rank test. Inter-rater consistency was evaluated using the intraclass correlation coefficient [ICC(2,1)] (two-way random, single measure), with ICC ≥ 0.75 considered indicative of good consistency. A *P*-value < 0.05 was considered statistically significant.

3. Results

3.1. Comparison of Model Response Efficiency and Long-Text Quality

As shown in **Table 1**, the average response time of GPT-5.5 was significantly shorter than that of Kimi K2.6 (9.2 ± 1.3 vs. 16.8 ± 2.5, $P < 0.001$), and the number of generated words was also lower than that of Kimi K2.6 (985 ± 156 vs. 1624 ± 287, $P < 0.001$). No statistically significant difference was observed in the overall scores between the two models (100.01 ± 1.45 vs. 100.30 ± 1.52, $P = 0.623$). The information density of GPT-5.5 was higher than that of Kimi K2.6 (9.85 ± 1.42 vs. 16.20 ± 2.95, $P < 0.001$). The information omission rate (4.2 ± 1.3 vs. 8.5 ± 2.1, $P < 0.001$) and cross-document consistency error (0.3 ± 0.2 vs. 1.2 ± 0.4, $P = 0.002$) of Kimi K2.6 were both significantly lower than those of GPT-5.5.

It is worth noting that although there is a statistically significant difference in response time between the two models (approximately 7.6 seconds), the practical impact of this time difference on teaching experience in real-world medical education scenarios is limited. Long-text teaching tasks—such as systematic lesson planning based on complete clinical guidelines and in-depth analysis of prolonged disease courses—are typically offline preprocessing steps; instructors can initiate model calls before class, and a difference of a few seconds in response time does not directly affect the fluency of classroom instruction. In contrast, quality metrics such as information omission rate and cross-document consistency error exert a more substantial influence on the accuracy and completeness of instructional content, and should therefore be considered core factors in model selection.

Table 1. Comparison of long-text teaching efficiency and quality between two large language models

Parameter	GPT-5.5 ($\bar{x} \pm s$)	Kimi K2.6 ($\bar{x} \pm s$)	*P value*
Mean response time (seconds)	9.2±1.3	16.8±2.5	<0.001
Generated text word count (words)	985±156	1624±287	<0.001
Total response quality score (points)	100.01±1.45	100.30±1.52	0.623
Word count-to-score ratio (words/point)	9.85±1.42	16.20±2.95	<0.001
Information omission rate (%)	8.5±2.1	4.2±1.3	<0.001
Cross-document consistency errors (times/task)	1.2±0.4	0.3±0.2	0.002
Context hallucination rate (times/task)	0.8±0.3	0.5±0.2	0.089

3.2. Comparison of Multi-Dimensional Score Distributions.

Figure 1 shows the distributions of the two models' ratings across six dimensions. GPT-5.5 outperformed

Kimi K2.6 in medical accuracy (17.11 ± 0.32 vs. 16.89 ± 0.28, $P = 0.089$), language adaptability (16.71 ± 0.29 vs. 16.15 ± 0.41, $P = 0.018$), and educational inspiration (17.25 ± 0.31 vs. 15.58 ± 0.22, $P < 0.001$); Kimi K2.6 outperformed GPT-5.5 in teaching logic (17.05 ± 0.30 vs. 16.47 ± 0.28, $P = 0.042$), long-text information integration ability (17.42 ± 0.33 vs. 15.83 ± 0.35, $P < 0.001$), and completeness of responses (17.21 ± 0.35 vs. 16.64 ± 0.38, $P = 0.003$).

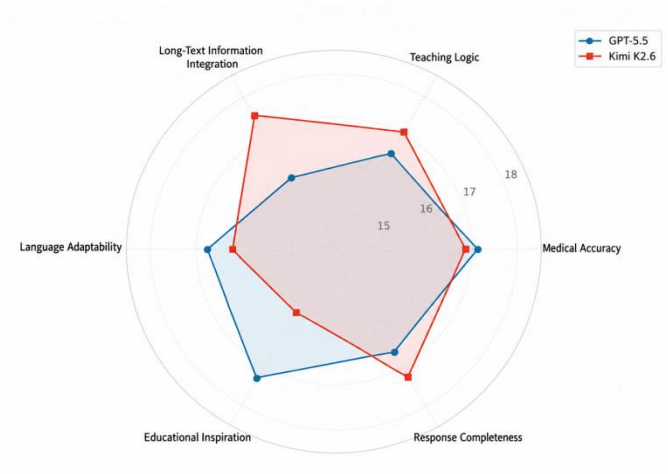

Figure 1. Radar chart of six dimension scores for GPT5.5 and Kimi K2.6 on long-text teaching tasks.

3.3. Analysis of score differences across task types

As shown in **Table 2**, for T1 tasks, the difference in total scores between the two models was not statistically significant ($P = 0.312$). Kimi K2.6 demonstrated superiority in long-text information integration capability ($P < 0.001$) and completeness of response ($P = 0.002$), while GPT5.5 outperformed in pedagogical inspiration ($P < 0.001$). For T2 tasks, Kimi K2.6 achieved a higher total score than GPT5.5 (101.3 ± 1.4 vs. 98.8 ± 1.6, $P = 0.008$), with particularly significant advantages in long-text information integration capability ($P < 0.001$) and teaching logic ($P = 0.004$). For T3 tasks, GPT5.5 scored higher than Kimi K2.6 (102.1 ± 1.3 vs. 97.5 ± 1.5, $P < 0.001$), with advantages concentrated in pedagogical inspiration ($P < 0.001$) and linguistic adaptability ($P = 0.011$). For T4 tasks, Kimi K2.6 significantly outperformed GPT5.5 in total score (103.5 ± 1.2 vs. 96.2 ± 1.8, $P < 0.001$), with the most prominent advantages in long-text information integration capability ($P < 0.001$) and cross-document consistency control ($P < 0.001$).

Table 2. Scores of the two large language models across different long-text teaching task types

Dimension	GPT5.5 ($\bar{x}$ ± s)	Kimi K2.6 ($\bar{x}$ ± s)	*P value*
T1 Guideline-based instructional design			
Total score	98.5±1.8	99.2±1.6	0.312
Long-text information integration	16.2±0.5	17.8±0.4	<0.001
Educational inspiration	17.0±0.3	15.2±0.4	<0.001
Response completeness	16.3±0.4	17.5±0.3	0.002
T2 Long case analysis			
Total score	98.8±1.6	101.3±1.4	0.008
Long-text information integration	15.5±0.4	18.0±0.3	<0.001

Dimension	GPT5.5 ($\bar{x}$ ± s)	Kimi K2.6 ($\bar{x}$ ± s)	*P value*
Teaching logic	16.2±0.4	17.5±0.3	0.004
T3 Multi-round Q&A			
Total score	102.1±1.3	97.5±1.5	<0.001
Educational inspiration	18.2±0.3	15.0±0.3	<0.001
Language adaptability	17.3±0.3	16.1±0.4	0.011
T4 Crossdocument integration			
Total score	96.2±1.8	103.5±1.2	<0.001
Long-text information integration	14.8±0.5	18.5±0.3	<0.001

3.4. Long-Text Specific Metric Analysis

As shown in **Table 3**, the overall information omission rate of Kimi K2.6 was significantly lower than that of GPT5.5 (4.2 ± 1.3 vs. 8.5 ± 2.1, $P < 0.001$). When stratified by document position, GPT5.5 exhibited significantly higher omission rates than Kimi K2.6 in the middle third (12.5 ± 3.2 vs. 5.1 ± 1.8, $P < 0.001$) and the last third (9.8 ± 2.5 vs. 4.7 ± 1.5, $P < 0.001$) of the documents, demonstrating a “Lost in the Middle” phenomenon; no statistically significant difference was observed between the two models in the first third ($P = 0.312$). In terms of cross-document consistency error, Kimi K2.6 outperformed GPT5.5 in tasks T2, T3, and T4, with the most pronounced difference in task T4 (0.8 ± 0.3 vs. 2.3 ± 0.6, $P < 0.001$); for task T1, which involved single-document input, both models yielded low error rates and the difference was not statistically significant ($P = 0.245$). Regarding contextual hallucination rate, the overall difference between the two models was not statistically significant (0.8 ± 0.3 vs. 0.5 ± 0.2, $P = 0.089$); however, GPT5.5 showed a trend toward higher hallucination rates in task T4 (1.2 ± 0.4 vs. 0.6 ± 0.2, $P = 0.052$).

Table 3. Comparison of long-text specific metrics between the two large language models

Stratification/Task Type	GPT5.5 ($\bar{x}$ ± s)	Kimi K2.6 ($\bar{x}$ ± s)	*P value*
Information omission rate (%)			
First 1/3 of document	3.2±1.1	2.8±0.9	0.312
Middle 1/3	12.5±3.2	5.1±1.8	<0.001
Last 1/3	9.8±2.5	4.7±1.5	<0.001
Overall	8.5±2.1	4.2±1.3	<0.001
Cross-document consistency error (times/task)			
T1	0.2±0.1	0.1±0.1	0.245
T2	0.8±0.3	0.2±0.1	0.008
T3	0.5±0.2	0.1±0.1	0.015
T4	2.3±0.6	0.8±0.3	<0.001
Contextual hallucination rate (times/task)			
T1	0.3±0.2	0.2±0.1	0.477
T2	0.5±0.2	0.3±0.2	0.245
T3	0.6±0.3	0.4±0.2	0.312
T4	1.2±0.4	0.6±0.2	0.052
Overall	0.8±0.3	0.5±0.2	0.089

4. Discussion

4.1. Core Findings and Pedagogical Interpretation

The results of this study indicate that the two representative large language models exhibit comparable overall teaching capabilities, yet show significant divergence across specific competency dimensions. This pattern—convergent total scores but divergent dimensional performance—suggests that current state-of-the-art LLMs have already moved beyond the stage of "whether they can be used" in medical education and entered a phase of refined adaptation focused on "how to select them according to specific needs." The differentiation in competency dimensions between models based on different technical pathways essentially reflects the diverse demands of long-text teaching tasks: while it is necessary to maintain faithful fidelity to original sources and simultaneously provide cognitive guidance to stimulate students' reasoning, a single model still struggles to fulfill both types of requirements simultaneously.

In terms of long-text information integration capability, the model with strong longcontext processing performed notably well in Tasks T2 and T4, with an information omission rate only half that of the comparison model, while the latter exhibited a typical "Lost in the Middle" phenomenon in the middle and later sections of documents. In teaching long-course case studies in hematology, the later parts of medical records often contain critical turningpoint information such as relapse/refractoriness and salvage therapy; omissions of such information can directly lead students to develop cognitive biases regarding incomplete disease natural history. The ability to retain distant context enables this type of model to maintain temporal coherence in longitudinal tracking teaching, which is of great value in hematology education for fully presenting the complete disease course from initial diagnosis to relapse/refractoriness for conditions such as leukemia and lymphoma.

However, some models showed an "over-integration" tendency in Task T4; when handling discrepancies between the WHO and ICC classifications for myelodysplastic syndromes, they tended to describe differences as "essentially consistent" rather than strictly presenting the numerical disparities in the original texts. While this "smoothing" approach improved response completeness, it may weaken the development of students' sensitivity to academic controversies. In contrast, other models were more inclined to directly state "Document A states..., Document B states..."; this "exposure of contradictions" strategy may be more valuable for cultivating critical thinking. This finding suggests that in cross-version document comparison teaching, the "depth of integration" in model outputs should align with instructional objectives—if the goal is to help students identify academic disputes, then "exposing contradictions" is preferable to "smoothing over differences."

In terms of educational heuristic quality, the model with strong chainofreasoning generation capability showed significant advantages in Task T3, using hypothetical counterquestions more frequently. It could construct counterfactual scenarios targeting specific logical gaps in students' reasoning, guiding them to self-discover cognitive biases. On the other hand, some models occasionally exhibited "context drift" during multi-turn dialogues, indicating that an ultra-long context window is not an absolute advantage for maintaining conversational coherence; there may be a trade-off between educational heuristic quality and long-text integration capability.

4.2. Analysis of Teaching Scenario Suitability

In addition to the traditional sixdimension scoring framework, this study added two long-text specific indicators to assess model fidelity to original teaching materials: the information omission rate reflects the

model's completeness in extracting key knowledge points from input documents—particularly in longcourse records or distal sections of guidelines, where omissions directly equate to missing teaching content; and the cross-document consistency error reflects the model's ability to identify and faithfully present version differences and contradictory statements when integrating multiple sources; excessive error may lead students to receive "smoothed-over" erroneous knowledge. Based on these indicators and the sixdimension scores, the differentiated positioning of the two models in digital hematology teaching can be further clarified.

In systematic lesson preparation and long-course case teaching, the model with strong long-text integration capability exhibited a lower information omission rate and could accommodate inputs of over 15,000 words without significant truncation; its generated teaching plans demonstrated higher structural completeness and more accurate citation of original guideline texts, making it suitable as an "intelligent lessonpreparation assistant." In clinical reasoning training and multi-turn Q&A sessions, the model with high educational heuristic quality stood out in terms of Socratic questioning and errorcorrection capability; its informationdensity advantage means that within limited classroom time, it can deliver equivalent teaching content in a more concise format.

For cross-version literature update comparisons, the two models showed task-specific differences in cross-document consistency error: in Task T1 (singledocument), both models yielded low errors (0.2±0.1 vs. 0.1±0.1, $P = 0.245$); in Tasks T2 and T3, the model with strong integration capability had lower errors; in Task T4, the difference was most pronounced (0.8±0.3 vs. 2.3±0.6 per task, $P < 0.001$). Accordingly, we recommend a collaborative model combining a "long-text integration model" and an "educational heuristic model": the former first performs preliminary difference extraction and table synthesis, and the latter then designs classroom discussion questions and cognitiveconflict scenarios based on the identified differences, thereby balancing breadth of information and depth of thinking.

4.3. Risk Prevention and Quality Control Recommendations

Although the overall hallucination rates of both models were low, the tendency of some models to fabricate updated guideline clauses in Task T4 suggests that when required to integrate multiple documents with version discrepancies, they may generate statements not present in the original texts in an attempt to "fill" logical gaps. In medical education, such hallucinations are particularly harmful because students often regard model outputs as authoritative references and may not detect fabricated content [18]. Therefore, regardless of the model chosen, all generated teaching materials must be verified by supervising instructors against original sources for critical knowledge points. Particularly when quantitative standards such as drug indications or diagnostic thresholds are involved, a three-tier quality control process—"model generation → source verification → instructor review"—should be established [19-21].

4.4. Limitations and Future Directions

Limitations of this study include: a limited number of tasks (24) without accounting for stochasticity in model outputs; comparison of only two models, which restricts generalizability; potential bias in expert ratings on subjective dimensions such as "educational heuristic quality"; and lack of validation of actual student benefits in real classroom settings. Future research should expand task scales, incorporate students as evaluation subjects, conduct controlled realclassroom trials, and explore hybrid architectures combining retrieval-augmented generation with longcontext models to constrain hallucination tendencies while

preserving long-text integration advantages.

5. Conclusions

This study indicates that large language models exhibit a pattern of "overall convergence with dimensional divergence" in long-text teaching scenarios within the Department of Hematology. Large language models based on different technical approaches have different emphases in capability dimensions: models with strong long-text integration capability are suitable for systematic lesson planning based on complete clinical guidelines and progressive teaching of longcourse cases, as they can effectively reduce information omission rates and ensure the completeness and temporal coherence of teaching content; models with high educational heuristic value are more appropriate for clinical reasoning training, multi-round Socratic questioning, and real-time classroom interaction, as they can stimulate students' critical thinking by making reasoning chains explicit and constructing counterfactual scenarios.

From the perspective of medical education practice, the introduction of large language models is not a simple replacement of technological tools, but rather a targeted response to the structural deficiencies of the traditional long-text teaching model. In the specialized field of hematology, which is highly dependent on standardized diagnostic and treatment pathways and long-term disease course tracking, model-assisted teaching can overcome the dual constraints of human working memory capacity and class time, thereby achieving a pedagogical paradigm shift from "fragmented knowledge transmission" to "systematic knowledge construction." However, regardless of how excellently a model performs in long-text integration or educational inspiration, its outputs still carry risks of information hallucination and excessive integration, and cannot replace the professional judgment and humanistic care of supervising teachers.

Therefore, when applying large language models in medical education, a three-tier application framework of "educationdemanddriven, modelcapabilitymatched, and teacherqualityassured" should be established: first, clarify the required capability dimensions (informationfidelityoriented or heuristicguidanceoriented) according to teaching objectives; then select the model type that matches that dimension; and finally, ensure the credibility of teaching materials through a three-level quality control process of "model generation, source verification, and teacher review." In the future, a multidimensional evaluation system covering information fidelity, cognitive scaffolding, and academic integrity should be further established to promote AI-assisted teaching from "usable" to "trustworthy" and "optimized use," thereby truly realizing the deep value of technology-empowered medical education

Funding

National Natural Science Foundation of China (Project No.: 82504923); Key Research and Development Program of Shaanxi Province (Project No.: 2022SF-239)

Disclosure statement

The author declares no conflict of interest.

References

[1] Zhang ZQ, Qi YH, 2025, Application status, challenges, and prospects of artificial intelligence large language models in the medical field. China Development, 25(6): 59-64.

[2] Yao Q, Wang Z, Guo YJ, et al., 2026, Principle of Large Language Model and Its Applications in Healthcare. Computer Systems and Applications, 35(1): 102-116.

[3] Wang YN, Jiang ZX, He HY, et al., 2025, Evolution of large language models and their applications in clinical medical education. West China Medical Publishers, 40(5): 777-782.

[4] Tong B, Tian C, Li YM, et al., 2026, Feasibility of Large Language Models Assisting in the Creation of Health Science Popularization Works. Medical Journal of Peking Union Medical College Hospital, 17(4): 1063-1072.

[5] Li Y, Li Z, Zhang K, et al., 2023, ChatDoctor: A Medical Chat Model Fine-Tuned on a Large Language Model Meta-AI (LLaMA) Using Medical Domain Knowledge. Cureus, 15(6): e40895.

[6] Umucu E, Solis G, Garza L, et al., 2025, Empathy by Design: Aligning Large Language Models for Healthcare Dialogue. 2025 IEEE International Conference on Big Data (BigData), 4693-4702.

[7] Tang CX, Pang QG, 2026, How LLMS and MCP reshape the future of medical education. China Medical Education Technology, 40(2): 177-186.

[8] Chen SF, Alyakin A, Seas A, et al., 2026, LLM-assisted systematic review of large language models in clinical medicine. Nature Medicine, 32(3): 1152-1159.

[9] Yin MZ, Li Q, Xiong WF, et al., 2020, A preliminary study on the application of PBL combined with CBL under multidisciplinary team collaboration mode in geriatric medicine teaching. Continuing Medical Education, 34(3): 22-24.

[10] Zhang LL, 2015, Application of PBL combined with CBL in clinical teaching of metabolic diseases. Journal of Modern Medicine and Health, (3): 461-463.

[11] Sun L, Liu XR, Lu XS, 2020, Problems and countermeasures of case-based teaching in early basic medical education. Basic Medical Education, 22(5): 327-329.

[12] Mennin S, 2021, Ten Global Challenges in Medical Education: Wicked Issues and Options for Action. Medical Science Educator, 31(1): 17-20.

[13] Eysenbach G, 2023, The Role of ChatGPT, Generative Language Models, and Artificial Intelligence in Medical Education: A Conversation With ChatGPT and a Call for Papers. JMIR Med Educ, 9: e46885.

[14] Shi Y, Yu K, Dong Y, et al., 2026, Large language models in education: a systematic review of empirical applications, benefits, and challenges. Computers and Education: Artificial Intelligence, 10: 100529.

[15] Qu X, Yang JM, Chen T, et al., 2023, Reflections on the Implications of the Developments in ChatGPT for Changes in Medical Education Models. Journal of Sichuan University (Medical Sciences), 54(5): 937-940.

[16] Wei MH, Liang JP, Li ZY, et al., 2026, A comparative study on the multidimensional performance of GPT-4o and DeepSeek-R1 in medical question answering in Hematology. China Medical Education Technology, 40(2): 222-227.

[17] Kimi K2: Open Agentic Intelligence. visited on May 5, 2026, https://arxiv.org/abs/2507.20534.

[18] Yang YY, Nie ZW, Zhou ZQ, 2026, Research on the Information Hallucination in Large Language Model Applications for Medical Education and Its Mitigation Strategies. Medical Education Research and Practice, 34(2): 293-302.

[19] Lin YH, Ye ZC, Chen YR, et al., 2026, Analysis of Dilemmas in the Integration of Generative Artificial Intelligence into Medical Education from the Perspective of Symbiotic Agency Theory. Medicine and Society,

39(3): 114-122.

[20] Ding XM, 2026, Research Progress on the Application of ChatGPT Technology in Medical Education. China Health Industry, 23(5): 128-131.

[21] Cong S, Bai WH, Chen Z, et al., 2026, Research Progress of Large Language Models in Case Generation, Personalized Teaching and Assessment. Journal of Medical Molecular Biology, 23(1): 97-106.