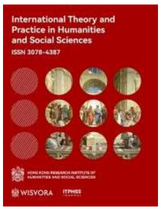
WISVORA

International Theory and Practice in Humanities and Social Sciences
2025 Volume2,Issue6 ISSN 3078-4387



Research on Input Characteristics in DeepSeek-Generated English Reading Comprehension Materials for China's National College Entrance Examination

Liu Qiang

College of International Education, Sichuan International Studies University, Shapingba, Chongqing,China

Accepted

2025-05-30

Keywords

AI-generated educational assessment,
input discourse characteristics,
Prompt engineering frameworks

Corresponding Author

Liu Qiang

doi.org/10.70693/itphss.v2i6.1056

Abstract

Under the digital transformation of educational evaluation in China, AI-assisted assessment has garnered unprecedented attention. This study focuses on the discourse features of English reading comprehension materials generated by DeepSeek for the National College Entrance Examination (Gaokao), employing a hybrid approach that integrates scientific prompt engineering frameworks with official documentation guidelines. The findings reveal that DeepSeek demonstrates limited proficiency in controlling micro-level textual features (e.g., length, lexical difficulty) but excels in managing macro-level features such as thematic relevance and genre alignment. This research contributes to advancing the digital transformation of educational evaluation in China, offering empirical data and methodological innovations for AI-generated assessment materials.

1. Introduction

Amid the global wave of digital transformation in education, China is advancing the innovation of its educational assessment system through a "dual-drive mechanism" synergizing policy guidance and technological empowerment. The 20th National Congress of the Communist Party of China explicitly outlined the strategic objective to "comprehensively enhance the quality of self-directed talent cultivation." Further reinforcing this vision, the Deepening Curriculum and Teaching Reform Action Plan for Basic Education (2023) issued by the Ministry of Education positions "digitally empowering teaching quality improvement" as a critical pathway. Within this context, the National College Entrance Examination (Gaokao), serving as the cornerstone mechanism for national talent selection, has seen the optimization of its question design quality and efficiency emerge as pivotal breakthroughs in educational assessment reform. Traditional English Gaokao question development relies on an "expert-document model" dominated by manual expertise, which suffers from prolonged timelines, high costs, and uneven coverage of assessed competencies. Such limitations render it increasingly inadequate to meet the refined demands of "New Curriculum Standards and the reformed Gaokao" for core literacy evaluation.

Meanwhile, breakthroughs in generative artificial intelligence (GenAI) technology have unlocked novel possibilities for educational assessment innovation. Represented by the domestically developed large language model DeepSeek, which integrates cross-modal semantic comprehension and pedagogical measurement knowledge embedding, this technology has pioneered the intelligent generation of standardized test items directly from textual sources. It not only rapidly parses the question design specifications outlined in the General Senior High School English Curriculum Standards but also optimizes the cognitive hierarchy distribution of test items through deep learning. According to statistics from the National Education Examinations Authority, the average cycle for developing English Gaokao questions in 2023 spanned 68 days, whereas AI-assisted systems can compress initial draft generation to under 5 hours—a 14-fold efficiency gain. Nevertheless, the academic community still lacks a systematic framework for validating the psychometric validity of AI-generated test items, particularly in critical dimensions such as textual input characteristics and cognitive skill mapping, which urgently require scientific evaluation.

Current research predominantly focuses on optimizing dynamic adjustment processes for AI-generated materials and test items, investigating methodologies to refine generation protocols for higher quality outputs and analyzing how different tuning approaches impact the content's quality and characteristics. However, there remains a paucity of studies examining the quality of standardized single-pass generation output, which means materials produced without iterative adjustments. While some scholars dismiss the value of studying first-attempt AI-generated content, this research posits that initial outputs are critical to subsequent quality optimization in AI-generated test items. Furthermore, prior studies have largely centered on small-scale, low-stakes assessments, whereas this study prioritizes large-scale, high-stakes examinations—specifically China's National College Entrance Examination (Gaokao) English tests. To address this research gap, this study employs a standardized instructional command framework to generate test materials and evaluates them using the "Task Characteristic Framework," particularly its discourse input feature dimensions. Specifically, it investigates the performance of AI-generated Gaokao reading comprehension materials across five parameters: length, reading speed, difficulty, themes, and genres. In other words, it explores how the discourse input features of DeepSeek-generated reading materials align with Gaokao requirements.

By doing so, this research aims to provide theoretical insights and empirical support for the digitization of educational assessment in China, while offering scholarly reference value for future studies in this field.

2. Literature Review

2.1Research on AI-Generated Reading Comprehension Test Items

Artificial Intelligence-Generated Reading Comprehension Test Items (Automatic Item Generation, AIG) refer to the automated process of generating reading comprehension questions that align with educational objectives and psychometric standards by leveraging natural language processing (NLP), cognitive modeling, and large language models (LLMs). Its core goals are to reduce manual item development costs, enhance efficiency, and ensure question diversity and quality. Unlike traditional human-authored items that rely on expert experience, AIG employs algorithmic models to analyze text semantics, extract key information, and generate questions and options through logical reasoning. For instance, template-driven methods populate predefined question templates with content, while NLP-driven approaches directly generate items using LLMs. (Jiang, 2025) proposed a technique for generating reading comprehension questions based on "key sentences" and question types, aiming to reduce dependency on answer keys and improve controllability.

Research on machine-generated reading comprehension tests has evolved through three phases globally: the Rule and Template Phase, Statistical and Machine Learning Phase, and Deep Learning and Generative AI Phase, with current studies predominantly focusing on the third stage. (Shin et al, 2023) integrated the structured features of templates with the flexibility of non-template methods by employing topic modelingto extract sub-topic information from texts, then generated inferential questions using prompt engineering to align with PIRLS standards. Their approach also filtered text difficulty via Lexile scores. (Sayin,Gierl,2024) utilized GPT-3.5 to generate distractors and design "irrelevant sentence identification" tasks, while (Lin,Chen , 2023) validated ChatGPT’s feasibility in producing multiple-choice questions, demonstrating comparable difficulty and discrimination indices to human-authored items. Comparative studies, such as (Shin et al, 2023), revealed that GPT-4-based Q-Craft outperformed traditional generators in question coherence and naturalness.

Current AI-generated reading comprehension models primarily include:Template-Driven Approach: Populating predefined templates with variables.NLP-Driven Approach: Directly generating items via LLMs.Hybrid Control Approach: Combining control mechanisms with generative models.

2.2Content validity

Content validity, as a core concept in educational and psychometric measurement, has undergone a paradigm shift from "comprehensive content coverage" to "multidimensional dynamic adaptability." Early definitions focused on the representativeness of test content to the target domain (Sireci, 1998), emphasizing whether items fully encompassed predefined knowledge or skill ranges (Chen Zhongyong, 1992). With advancements in validity theory, (Messick, 1989) proposed the unified validity framework, advocating for the integration of content validity with construct validity, social validity, and other dimensions. This perspective highlights the dynamic alignment of test content with real-world competency demands, cognitive processes, and pedagogical objectives, driving the evolution of content validity from "static content sampling" to "contextualized competency mapping."

In language testing, Bachman ’ s task characteristics model expanded the theoretical framework of content validity by incorporating "input characteristics", "expected responses" (e.g., cognitive skill requirements), and "task conditions" into the evaluation system. （Fulcher, 1999) introduced needs analysis, proposing that content validity verification must be grounded in task feature analysis of the Target Language Use (TLU) domain to ensure ecological consistency between test tasks and real-world language behaviors. Recent scholarship, exemplified by (Young, 2009), emphasizes cross-cultural fairness, arguing that content validity assessments must account for the impact of multicultural backgrounds on task accessibility. This marks the deepening of content validity into socioculturally sensitive evaluation.

By the late 1990s, empirical research began to emerge in China. Scholars like (Jin Yan, 1998) evaluated the content validity of specific exams, such as the College English Test (CET), using mixed methods (quantitative analysis and introspective techniques) to explore the alignment between test items and authentic reading behaviors. Globally, early studies focused on whether tests comprehensively covered predetermined content domains (Sireci, 1998). Subsequent research shifted toward ensuring representativeness and adaptability through content analysis and needs analysis, broadening the scope of content validity from mere content coverage to a holistic analysis of relationships among language competencies, test tasks, and assessment goals.

In the early 21st century, content validity definitions expanded further, particularly in language testing. Scholars integrated needs analysis and task analysis to align test content with practical requirements (Fulcher, 1999). (Young, 2009) framework refined content validity research by emphasizing not only task representativeness but also cross-group fairness for diverse student populations.

2.3Discourse Input in Language Testing

In language testing, "discourse input" refers to the linguistic materials and accompanying contextual features presented to test-takers, with its core function being to elicit specific cognitive behaviors for measuring target language abilities (Bachman & Palmer, 1996). Traditional definitions focused on static textual attributes such as lexical complexity, syntactic structures, and genre types. However, as validity theory evolved, its conceptual scope expanded to encompass dynamic interactive dimensions. Bachman and Palmer’s task characteristics model redefined discourse input as a three-dimensional construct comprising "input format" , "linguistic features" , and "contextualized information" . This framework emphasizes simulating real-world language use scenarios (Target Language Use Domain, TLU) to achieve ecological validity (Fulcher, 1999). Recent advancements propose a cognition-oriented perspective, advocating that discourse input design must align with test-takers’ information processing mechanisms (e.g., attention allocation, inferential pathways) and be empirically validated through methods like eye-tracking to assess its contribution to ability measurement (Young, 2009; Koreeda et al., 2021).

Discourse input studies have transitioned from a text-centric to a cognitive-social interaction paradigm. Early research (1980s–1990s) prioritized surface-level textual analysis, such as (Carrell, 1985) use of readability formulas to quantify text difficulty, though this approach overlooked context’s role in comprehension.

Post-1990s, (Bachman&Palmer, 1996) task characteristics theory shifted focus to input-response dynamics. For instance, (Fulcher, 1999) analyzed IELTS reading materials through the TLU lens, revealing that academic tests should prioritize authentic texts like journal abstracts.

In the 21st century, technological advancements spurred new directions. First, computerized testing introduced multimodal elements into discourse input design, though debates persist about their cognitive load effects (Chapelle & Douglas, 2006). Moreover, generative AI has redefined discourse input generation logic. While AI-generated texts approach human-authored levels in syntactic complexity, studies identify systematic deviations in semantic focus distribution (Wang et al., 2023).

Current research gaps center on validating AI-generated discourse inputs, particularly the absence of standardized frameworks to align them with curriculum-based competency mappings.

3. Methodology and Procedures

3.1Methodology

3.1.1Textual Analysis

In this study, researchers conducted textual analysis on reading materials generated by DeepSeek for China’ s National College Entrance Examination (Gaokao). The analysis focused on five dimensions: text length, reading speed, text difficulty, textual themes, and genre types.

3.1.2Delphi Method

The Delphi method is a systematic approach that gathers expert opinions through iterative consultations to reach consensus on specific research questions. In this study, a panel of language testing experts and practitioners will be engaged to evaluate AI-generated materials, addressing key questions such as:What genre does this reading material belong to? Why?What theme or subject matter does this reading comprehension material address? Why?

The expert panel comprises:

10 postgraduate students specializing in English education from Sichuan International Studies University (Chongqing, China)

1 professor who expertise in language assessment

1 professor who expertise in English education

1 professor who expertise in pedagogy

2 senior high school English teachers from provincial-level model schools

This diversified composition—spanning academia, pedagogical research, and frontline teaching—ensures methodological rigor and enhances the credibility of the study’s conclusions.

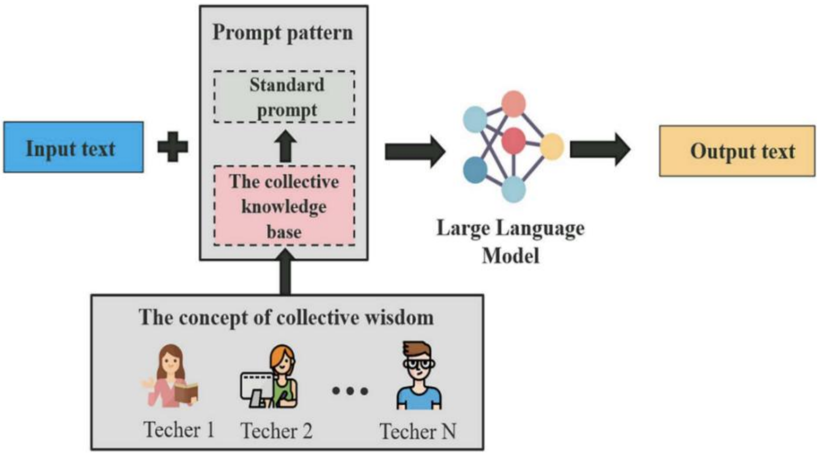
3.2Procedures

The study analyzed reading comprehension materials generated by DeepSeek, an AI-powered system. The research process mainly consists of two parts. The first part is to generate materials, and the second part is to analyze materials.

3.2.1Material Generation

Reading materials were generated using prompt engineering frameworks， proposed by （Wang Lili， 2023), which includes six parts(Role 、 Output Indicator, Type, Definition, Characteristic and Example Code) adhering to the *2019 Syllabus of National College Entrance Examination, the General Senior High School English Curriculum Standards*, and the China's Standards of English Language Ability, referencing the text length and difficulty levels of the *National New Curriculum Volume II* reading comprehension tasks.

Figure3.1(Wang Lili,2023)Model of AI-based items generation



3.2.2 Material Analysis

This study adapts the discourse input analysis framework（Table3.2） proposed by (Dong, 2010) under Bachman's task characteristics model, with modifications aligned to China's official General Senior High School English Curriculum Standards (2017).

Text Length: Measured via Microsoft Word's built-in word count function.

Reading Speed: Calculated using the formula *Total Words / Reading Time = Reading Speed* (words per minute).

Text Difficulty: Assessed using the "Language Data" computational tool developed by Jin Tan (2023).

Theme & Genre Classification: Determined through the Delphi method, where experts independently categorized materials and reached consensus through iterative feedback.

Table 3.2: discourse input analysis framework

Item	Description
Article Length	Single article length, Total length of multiple articles
Reading Speed	Reading speed of articles, Completion speed of entire reading comprehension
Article Difficulty	Number of new words, Readability
Theme	Human and Nature, Human and Society, Human and Self
Genre	Narrative, Expository, Argumentative, Practical

4. Results and Discussion

After generating reading materials and test questions using DeepSeek (DepthSeeker R-1 large model), textual and data analyses were conducted on the generated reading materials based on five dimensions of discourse input. The analysis results consist of three categories of data:

Instructional language data: Predefined values or ranges designed in instructional guidelines.

AI self-inspection data: Automated quality reports generated by DeepSeek itself after content generation, provided to users for validation.

Actual data values: Empirical measurements obtained by researchers through analytical tools in this study, used to identify discrepancies between expected and actual values across discourse input dimensions.

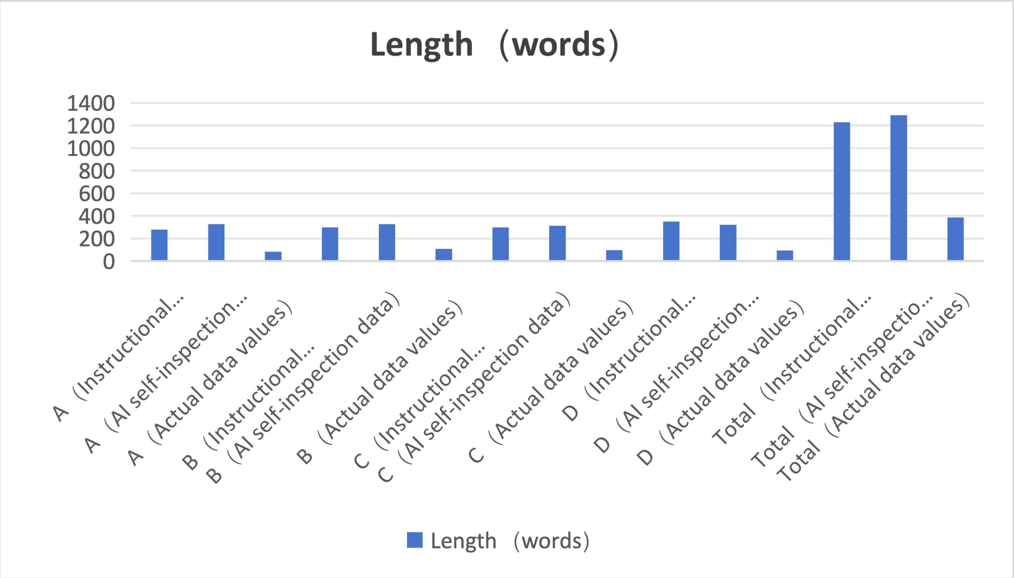
Table 4.1

Passage（ABCD）	Length（words）	Reading speed（words/ minutes）	Difficulty	Theme	Genre
A（Instructional language data）	280	40	4.7	human and society（environmental protection activity）	Practical
A（AI self-inspection data）	328	40	4.75	human and society（environmental protection activity）	Practical
A（Actual data values）	85	14	4.3	human and society（environmental protection activity）	Practical
B（Instructional language data）	300	40	4.8	human and self（ growth and insights）	narration interspersed with comments
B（AI self-inspection data）	328	40	4.75	human and self（ growth and insights）	narration interspersed with comments
B（Actual data values）	109	13.6	4.1	human and self（ growth and insights）	narration interspersed with comments
C（Instructional language data）	300	40	4.9	human and nature（ecological science）	Expository
C（AI self-inspection data）	312	40	4.8	human and nature（ecological science）	Expository
C（Actual data values）	97	12	4.2	human and nature（ecological science）	Expository
D（Instructional language data）	350	40	4.9	human and society（influence from technology）	Argumentative

D（AI self-inspection data）	323	40	4.8	human and society（influence from technology）	Argumentative
D（Actual data values）	95	11.8	4.3	human and society（influence from technology）	Argumentative
Total（Instructional language data）	1230	40	4.8		
Total（AI self-inspection data）	1291	44	4.7		
Total（Actual data values）	386	12	4.2		

4.1Length

Figure4.2



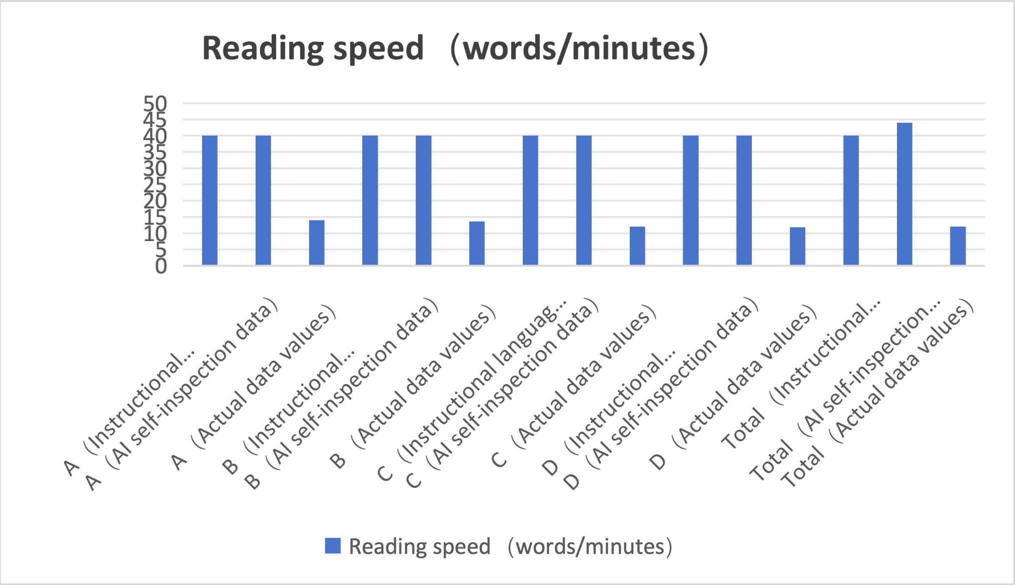
During the instruction editing process, the predefined word count targets were set as follows: Passage A (280 words), Passage B (300), Passage C (300), and Passage D (350). After generating the reading materials and test questions, DeepSeek's self-inspection report showed actual word counts of 328, 328, 312, and 323 respectively for the four passages. The maximum single-passage discrepancy reached 48 words (17% variance rate), while the minimum gap was 12 words (4% variance). Overall, the total target word count of 1,230 words across all four passages showed a 61-word discrepancy (4.9% variance) compared to the self-reported 1,291 words. Historically, the total word count of reading comprehension materials in China's National College Entrance Examination (NCEE) English Paper II over the past three years has consistently ranged between 1,200-1,400 words. Therefore, based on instructional specifications and AI self-inspection data, DeepSeek's word count generation demonstrates acceptable compliance despite minor variances.

However, substantial discrepancies emerged between actual measurements and instructional specifications:Passage A: 85 words (≈3× shorter than target)， Passage B: 109 words (191 words below target)， Passage C: 97 words (193 words below target)， Passage D: 95 words (255 words below target).The total actual word count of 386 words showed a 844-word deficit (68% variance rate) compared to the 1,230-word target. This performance fails to meet both expected standards and NCEE requirements.

Potential causes for this mismatch between reported and actual word counts include:Tokenization Mechanism,LLMs generate text based on tokens rather than characters/words, Chinese character-to-token ratio (≈1:1.5) may cause conversion errors, and possible algorithmic bias in API word count calculations due to improper reverse conversion. In addition, Generation Termination, early termination triggered by stop conditions, mechanical output of predefined completion statements despite premature termination. Finally, initial-phase direct generation without iterative instruction optimization, which causes lack of dynamic adjustment mechanisms for text length and quality control.

4.2Reading speed

Figure4.3



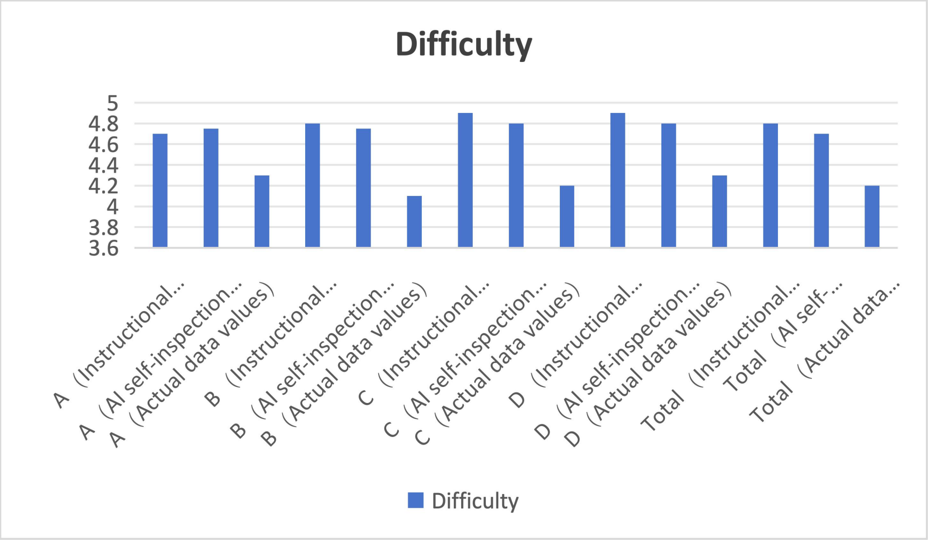
Reading speed is determined by the total word count and recommended reading duration. This study conducted a preliminary analysis of the National Curriculum Standards (NCS) English Paper II examinations (2022-2024), revealing that all four reading comprehension passages consistently prescribed a 30-minute reading timeframe. Consequently, actual reading speed was calculated as actual word count divided by reading time. Given the variations in passage length and question quantity – specifically, shorter passages A-B (with A containing 3 multiple-choice questions versus 4 questions each for B, C, and D) – the recommended time allocation per passage was respectively 6, 7, 8, and 9 minutes. As demonstrated in the Table 4, the comparison between instructional specifications and AI self-inspection data for the four passages showed minimal discrepancies and met basic compliance standards.

However, empirical measurements revealed critical deficiencies in reading speed performance. The actual reading speeds for passages A-D measured 14 wpm, 13.6 wpm, 12 wpm, and 11.8 wpm respectively, yielding an aggregate speed of 12 wpm – substantially below the three-year NCS Paper II average of 40 wpm. This indicates that DeepSeek's directly generated reading materials failed to meet expectations and examination requirements regarding reading speed benchmarks.

The primary cause of this inadequacy stems from insufficient text length generation. The system-produced passages (A: 85 words, B: 109 words, C: 97 words, D: 95 words) exhibited severe truncation compared to specifications (A: 280 words, B: 300 words, C: 300 words, D: 350 words), resulting in disproportionately short texts that disrupted normal reading rhythm and compromised time allocation mechanisms essential for examination simulations.

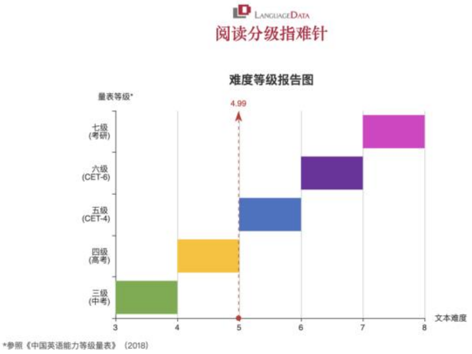
4.3Difficulty

Figure4.4



For text difficulty equivalence classification, the Languagedata Reading Difficulty Compass was employed. As illustrated (Figure4.5), analysis of Passage D from the 2024 National New Curriculum Standards English Paper II demonstrated a difficulty score of 4.99, indicating alignment with the gaokao-level difficulty benchmark and surpassing 99% of comparable reading materials in complexity within its category.

Figure4.5



For the difficulty equivalence classification of reading comprehension materials, this study adopted the *Languagedata Reading Difficulty Compass*（Jintan，*2023*）. Over the past three years (2022-2024), the average difficulty level of reading comprehension passages in China's National New Curriculum Standards English Paper II was 4.86. Accordingly, instructional specifications for generating four passages (A-D) were set at 7.7, 4.8, 4.9, and 4.9 respectively. As shown in

Figure 4.3, the AI self-inspection data revealed that Passage A achieved a difficulty score of 4.75 (exceeding the target of 4.7), while the other three passages exhibited marginally lower values than their respective targets, yielding an overall difficulty of 4.7 that met basic gaokao requirements. However, empirical measurements demonstrated significantly reduced difficulty levels: individual passage scores of 4.3, 4.1, 4.2, and 4.3, with a composite difficulty of 4.2 – substantially below the three-year examination average yet aligning with Languagedata's Level 4 criteria.

The observed discrepancies between generated material difficulty and instructional specifications stem from four primary factors:

First， training Data Limitations, the model was trained on only one year's gaokao authentic materials due to computational constraints in processing extensive examination datasets. This restricted exposure impaired the AI's ability to internalize examination-specific lexical complexity, syntactic sophistication, and stylistic patterns critical for difficulty control.

Furthermore， architectural Constraints, the probabilistic generation nature of Transformer models prioritizes text coherence over deliberate complexity crafting. Local attention mechanisms favor high-frequency linguistic patterns, conflicting with the intentionally designed low-frequency complex structures characteristic of gaokao texts.

Additionally, control Signal Attenuation, abstract instructional cues like "gaokao-level difficulty" exhibit progressive signal decay during generation. Later portions of texts demonstrate weakened adherence to difficulty constraints compared to initial segments, compromising overall complexity management.

Finally, dynamic Adjustment Deficiency, the direct generation mode lacks iterative optimization mechanisms. Without real-time calibration based on generated text features or corpus re-tuning, the system cannot dynamically reinforce difficulty control parameters during the generation process.

These technical limitations collectively account for the systemic challenges in replicating examination-authentic text complexity through current AI generation paradigms.

4.4Theme and Genre

The performance of DeepSeek in generating reading comprehension materials demonstrated acceptable compliance in thematic content and text types. As shown in Table 4.1, the instructional specifications, AI self-inspection reports, and empirical measurements aligned consistently across three data categories, meeting the diversified thematic requirements of "Human-Society, Human-Nature, and Human-Self" outlined in gaokao assessment standards. Regarding text types, the generated materials included narratives, expository texts, blended narrative-commentary, and argumentative essays, conforming to the 2019 National English Gaokao Outline's requirements for textual genre diversity.

The observed dichotomy – effective genre/thematic control versus inadequate length/difficulty regulation – stems from hierarchical differences in text feature controllability and inherent characteristics of model architecture, explained through four analytical dimensions:

4.4.1Explicit Imitability of Structural Features

Thematic and generic features exhibit distinct patternization. Gaokao-standard argumentative/expository texts follow stable discourse structures (e.g., thesis-support-conclusion frameworks), while recurring themes like technological ethics or cultural heritage correspond to specific keyword networks ("AI", "intangible cultural heritage"). These discrete categorical patterns, reinforced through high-frequency occurrences in training data, create strong statistical signals that models effectively capture through attention mechanisms.

4.4.2Continuous Control Paradigms

Genre/thematic selection constitutes finite-set discrete choices analogous to multi-class classification, where models optimize probability distributions. Conversely, text length/difficulty regulation requires continuous spectrum control through dynamic adjustment of micro-features, like sentence length variance and lexical difficulty gradients. Current models lack precise quantitative feedback mechanisms to maintain stability in these granular parameters during generation.

4.4.3Asymmetric Data Annotation

Pretraining corpora inherently contain abundant genre labels, including news and academic papers, thematic keywords, including environment and education, providing clear learning objectives. However, text difficulty as an implicit feature lacks standardized annotation protocols and relies on expert judgment, such as Flesch readability indices, hindering reliable difficulty prediction frameworks.

4.4.4Cognitive Resource Allocation Priorities

Transformer architectures prioritize high-level attention patterns for macro-structural framing during initial generation phases. Difficulty control necessitates continuous monitoring of micro-features, including lexical selection and syntactic complexity, requiring computationally intensive fine-grained adjustments. The decoding process exhibits "feature attenuation" – progressively weakening adherence to initial difficulty constraints as text generation proceeds.

This phenomenon reveals an asymmetry in generative AI's capabilities between pattern recognition and parameter fine-tuning, demonstrating that current technology excels at handling explicit, discrete structural features while exhibiting limitations in managing implicit, continuous complexity controls. For educational AI applications requiring strict adherence to pedagogical standards, this suggests the necessity for hierarchical control systems that decouple macro-structural generation from micro-level difficulty calibration.

5. Conclusion and Suggestion

5.1Conlusion

This study utilizes a scientific instructional command structure (Wanglili, 2023) and relevant official document descriptors, including the English Curriculum Standards for Senior High Schools (2017 Edition), the 2019 National College Entrance Examination English Syllabus, and China’s Standards of English Language Ability, to generate reading comprehension materials for China’s National College Entrance Examination (Gaokao) English tests through the

DeepSeek-R1 large language model. The findings indicate that AI-generated Gaokao English reading comprehension materials underperform in discourse representation. Specifically, their fine-grained control over discourse length and difficulty fails to align with the average standards observed in the Gaokao English test papers from the preceding three years. However, the materials exhibit unexpectedly strong performance in thematic and genre control, demonstrating the ability to accurately generate target topics and text types that fully comply with predefined instructional requirements.

Consequently, this study concludes that Gaokao English reading comprehension materials generated via DeepSeek, standardized instructional frameworks, and official document descriptors perform inadequately in micro-level control, falling entirely short of expectations, while excelling in macro-level control. These outcomes are attributed to three primary factors:

AI Text Generation Mechanisms: Limitations inherent to DeepSeek, such as discrepancies in token counting across languages and inconsistent adherence to instructional constraints during generation.

Experimental Design: The study's protocol prioritized rigorous pre-experiment preparations (e.g., structured instructional commands and official document descriptors) but omitted dynamic adjustments based on initial output quality, relying instead on post-generation data analysis and expert evaluation.

High-Stakes Testing Standards: The stringent quality benchmarks of China's Gaokao—a high-stakes, large-scale examination—pose significant challenges for AI-generated materials to meet required proficiency levels in their first iteration.

This research underscores the current limitations of AI in producing high-stakes educational assessment materials while highlighting its potential for thematic and genre adaptability under structured guidance.

5.2Suggestion

The findings of this study highlight both the potential and limitations of generative AI in high-stakes assessment design. While the instructional framework proposed by Wang (2023) has demonstrated empirical validity in general educational question design, its direct applicability to high-stakes, large-scale examinations like the Gaokao remains constrained by domain-specific complexities. To address these challenges, future research should prioritize the following interdisciplinary advancements:

5.2.1Development of Specialized Prompt Engineering Frameworks

This study drew on the prompt mining machine of (wanglili, 2023), but scientific prompt mining machines, especially the prompt framework for the automatic generation of test questions, urgently need further exploration. Hence, a critical frontier lies in designing exam-centric prompt architectures that systematically operationalize official assessment criteria， such as China's Standards of English Language Ability， into machine-actionable parameters. This requires several elements. First, the hierarchical Decomposition, creating multi-layered prompt structures that disentangle macro-level requirements ,including genre, thematic alignment, from micro-level constraints, lexical difficulty bands, syntactic complexity thresholds. Second, Cross-Modal Alignment: Integrating linguistic corpora ,such as old Gaokao papers, with computational metrics to establish quantifiable mappings between human-defined standards and machine-executable rules. Finally, the domain adaptation, leveraging techniques from psychometric testing,like item response theory, to calibrate prompts for examination-specific validity, reliability, and fairness requirements.

5.2.2 Dynamic Feedback-Driven Generation Systems

This study adopted the initially generated reading materials for analysis. Future research can continuously explore the dynamic adjustment mechanism and improve the quality of test questions. The single-pass generation paradigm, as employed in this study, fundamentally limits AI's capacity to meet high-stakes quality thresholds. A paradigm shift toward iterative refinement systems could mitigate this. First, Closed-Loop Optimization, implementing reinforcement learning frameworks where initial outputs are evaluated against multidimensional quality metrics , with discrepancies triggering automatic prompt recalibration. Moreover, Human-in-the-Loop Hybridization, developing hybrid workflows where AI-generated drafts undergo expert validation, with annotated feedback being reinjected into subsequent generation cycles. Last,Real-Time Complexity Modulation, exploring transformer-based architectures capable of dynamically adjusting generation parameters during text production.

5.2.3Domain-Specific Calibration Protocols

To resolve the observed micro-level control deficiencies, future systems require assessment-aware calibration mechanisms.Language-Specific Tokenization Standards, redesigning token counting algorithms to accommodate cross-linguistic disparities, particularly for Chinese-English educational texts where mismatches in morphological complexity distort length control. Additionally, Granular Difficulty Scaffolding, implementing tiered difficulty parameters that synchronize with official proficiency scales , potentially through adversarial training with discriminator models trained on ranked examination corpora. Besides,Temporal Compliance Modules, architecting time-sensitive generation constraints that ensure alignment with evolving examination trends, utilizing techniques from diachronic corpus analysis to detect and adapt to stylistic shifts in historical test papers.

5.2.4 Validation Frameworks for Generative Assessment Systems

Establishing rigorous evaluation methodologies specific to AI-generated examination materials needs Psychometric-Grounded Metrics, moving beyond surface-level textual analysis to develop validity indices measuring construct representation and consequential validity. It also need Cross-Institutional Benchmarking, creating open-access repositories of AI-generated and human-authored test items to facilitate large-scale comparative studies on fairness, bias, and pedagogical utility.

By bridging computational innovation with psychometric rigor, such research trajectories could transform generative AI from a supplementary tool into a validated component of next-generation educational assessment ecosystems.

Acknowledgments

Sincere gratitude is extended to Sichuan International Studies University for its financial support of this research. The author deeply appreciate the expertise and constructive feedback provided by the participating scholars and practitioners. Finally, we extend our professional thanks to the reviewers and editors for their rigorous guidance and invaluable dedication throughout the publication process.

Funding

This study was supported by grant from Postgraduate Scientific Research Innovation Project of Sichuan International Studies University(SISU2024XK120).

References

Bachman, L. F., & Palmer, A. S. Language testing in practice. Oxford University Press. (1996):89-122.

Messick, Samuel. "Validity and washback in language testing." Language testing 13.3 (1996): 241-256.

Jiang, Y. R., Tao, Y. Y., Wang, X., et al. Research on Reading Comprehension Question Generation Technology Based on Key Sentences and Question Types [J/OL]. Computer Engineering and Applications, 2025: 1-17.

Lee, J. H., Shin, D., & Noh, W. (2023). Artificial Intelligence-Based Content Generator Technology for Young English-as-a-Foreign-Language Learners' Reading Enjoyment. RELC Journal, 54(2), 508-516.

Sireci, Stephen G. "The construct of content validity." Social indicators research 45 (1998): 83-117.

Chen, Z. Y. (1992). On the Validity Issues of Educational Testing [J]. Journal of Inner Mongolia Normal University (Philosophy and Social Sciences Edition), (03): 106-111.

Fulcher, Glenn. "Assessment in English for academic purposes: Putting content validity in its place." Applied linguistics 20.2 (1999): 221-236.

Young, J. W. (2009). A framework for test validity research on content assessments taken by English language learners. Educational Assessment, 14(3–4), 122–138.

Jin Yan. (1998). A study on the content validity of CET reading comprehension tests. Modern Foreign Languages, (2), 61–67.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), Educational Measurement (3rd ed., pp. 13–103). Macmillan.

Sireci, S. G. (1998). The construct of content validity. Social Indicators Research, 45(1–3), 83–117.

Carrell, P. L. (1985). Facilitating ESL reading by teaching text structure. TESOL Quarterly, 19(4), 727–752.

Chapelle, C. A., & Douglas, D. (2006). Assessing Language Through Computer Technology. Cambridge University Press.

Koreeda, A., et al. (2021). Cognitive validity in reading tests: A systematic review. Language Testing, 38(3), 456–480.

Wang, J., et al. (2023). Evaluating AI-generated reading materials: A psycholinguistic approach. Computer Assisted Language Learning, 36(5), 1023–1045.

Wang, Lili, et al. "Exploring prompt pattern for generative artificial intelligence in automatic question generation." Interactive Learning Environments (2024): 1-26.

Dong, M. X. (2011). A diachronic study on the content validity of English reading comprehension tests in Chongqing college entrance examinations (2004–2009). Educational Measurement and Evaluation (Theoretical Edition), (2), 52–56+22.

Ministry of Education of the People's Republic of China. English Curriculum Standards for Senior High Schools (2017 Edition). People's Education Press, 2017.