Huilin Chen, 2026

Volume 10 Issue 3, pp. 12-29

Received: 06th April 2026

Revised: 26th June 2026

Accepted: 03rd September 2026

Date of Publication: 25th November 2026

DOI- https://doi.org/10.20319/pijtel.2026.103.1229

This paper can be cited as: Chen, H. (2026). Expert Evaluation-Based Comparative Evaluation of Three LLMS for CEFR-Aligned Reading Passage and Item Generation. PUPIL: International Journal of Teaching, Education and Learning , 10(3), 12-29

This work is licensed under the Creative Commons Attribution-Non-commercial 4.0 International License. To view a copy of this license, visit http://creativecommons.org/licenses/by-nc/4.0/ or send a letter to Creative Commons, PO Box 1866, Mountain View, CA 94042, USA.

EXPERT EVALUATION-BASED COMPARATIVE EVALUATION OF THREE LLMS FOR CEFR-ALIGNED READING PASSAGE AND ITEM GENERATION

Huilin Chen
School of Education, Shanghai International Studies University, Shanghai, China
chlwilliam@hotmail.com

Abstract

This study systematically evaluates the automated item generation (AIG) capability of three large language models—Deepseek, GPT-4, and ERNIE Bot—across three key dimensions: text quality, skill targeting, and cross-proficiency stability. Based on expert ratings of 24 passages and 120 items across six CEFR levels (A1–C2), the results show that: (1) for text generation, Deepseek and ERNIE Bot significantly outperform GPT-4

in matching text difficulty to CEFR levels ($p<0.001$); (2) for skill-based generation, Deepseek demonstrates a significant advantage over ERNIE Bot in producing inference-type items ($p=0.022$); and (3) for cross-level stability, all models perform comparably from A1 to B2, but Deepseek significantly surpasses GPT-4 at the advanced C2 level ($p=0.002$). These findings provide empirical guidance for model selection in test development and highlight the growing competitiveness of domestic Chinese LLMs in high-stakes language assessment.

Keywords:
Large Language Models, Automated Item Generation, Expert Evaluation, Text Generation, CEFR

I. Introduction

With the rapid development of artificial intelligence technology, large language models have demonstrated unprecedented capabilities in the field of natural language processing, which is profoundly changing the research paradigm of language testing. As the core link in language test development, item quality directly affects the validity and fairness of the test. Automated item generation has long been an important direction in language testing research. The emergence of large language models offers new possibilities for automated item generation: they can quickly generate high-quality test items, thereby significantly shortening test development cycles, reducing labor costs, and laying the foundation for personalized testing.

However, current research on automated item generation is still in its preliminary stages, and there is almost no systematic research on the use of large language models for test item generation. Existing studies have mostly focused on the ability of models in text generation or language comprehension, lacking systematic research on their specific application value and reliability in language testing item development. Based on this, this study aims to focus on evaluating the ability of large language models in automated item generation, exploring their applicability across

different texts and skills through empirical analysis, and providing theoretical and practical references for future automated and intelligent language test development.

The innovations of this study are as follows: on the one hand, it systematically evaluates the actual performance of large language models in language testing item development, providing a scientific basis for quantifying model capabilities; on the other hand, it offers practical guidelines for the application of automated item generation in language assessment, with potential educational technology promotion value. Through an in-depth analysis of the item generation capabilities of large language models, this study not only enriches the research methods of language testing but also lays the foundation for the realization of intelligent and personalized language assessment.

II. Literature Review

2.1 Review of Automated Reading Item Generation

Automated item generation refers to the process of automatically generating items by machine according to specific requirements (Lin & Chen, 2024). With the rapid development of generative artificial intelligence technology, automated item generation has gradually become a research hotspot in the fields of language testing and educational technology. In recent years, researchers have conducted extensive explorations around different item types and application scenarios, including reading cloze items (Yang et al., 2021), multiple-choice question generation (Ahmed et al., 2024), automated item generation and difficulty prediction (Zhu et al., 2024), and listening material generation (Aryadoust et al., 2024). For example, Lin and Chen (2024) explored ChatGPT's ability to generate multiple-choice questions and found that the model can generate items based on specified reading skills, and the quality of the generated items is comparable to that of human-written items. Similarly, Aryadoust et al. (2024) studied the feasibility of generating listening materials based on large language models, and the results showed that the materials generated by the model had good discriminant validity. Although the above studies provide important references for

the application of large language models in automated item generation, there are still limitations. First, previous studies generated a small number of items and lacked empirical research on large-scale item generation. Second, items were mostly concentrated at a specific proficiency level and rarely covered different levels and topics. Therefore, it is necessary to systematically evaluate the generation ability of large language models under different proficiency levels and topic conditions to compensate for the shortcomings of existing research.

2.2 Comparison of Automated Item Generation Capabilities of Large Language Models

With the development of generative artificial intelligence, various large language models have emerged. The representative foreign classic model is the GPT series, such as GPT-4; domestic large language models include ERNIE Bot and Deepseek. These models provide diverse options for the research and application of automated item generation. However, systematic evaluation research on the automated item generation capabilities of large language models is still in its preliminary stage. For example, Liu et al. (2025) systematically compared the automated item generation capabilities of models such as GPT-3.5, GPT-4, Llama2, Llama3, Gemini, and Cohere. The study found that GPT-4 performed best, and the quality of its generated items was close to that of human-written items. However, this study did not include domestic large language models such as Deepseek, which has advantages in reasoning ability and is particularly suitable for generating inference-type items, which has important application potential in automated item generation. Therefore, it is necessary to systematically compare domestic and international mainstream large language models to evaluate their actual capabilities in text generation and skill-based generation, providing empirical references for the application of large language models in automated item generation.

2.3 Research Questions

To comprehensively evaluate the ability of large language models in

automated item generation tasks and to consider the practical application background in China, this study focuses on domestic large language models (such as ERNIE Bot and Deepseek) while also including the international mainstream model GPT-4. The specific research questions are as follows:

(1) How do different large language models perform in terms of text generation ability?

(2) How do different large language models perform in terms of generating items based on specified reading skills?

(3) How consistently do these models generate high-quality items across six CEFR proficiency levels, and do their performances differ significantly at specific advanced levels?

III. Research Methods

3.1 Model Selection

This study selected three large language models, including GPT-4, Deepseek, and ERNIE Bot. The reasons for selecting these three models are that they are all representative language models in recent years, with GPT-4 being an internationally renowned model, and Deepseek and ERNIE Bot being domestic models. By including both domestic and international representative models, this study can more comprehensively evaluate the capabilities of different types of large language models in automated item generation, providing a reference for the selection and application of language models.

3.2 Text Generation

Text is the core foundation of reading item development, so the first step in automated item generation is the generation of reading texts. Referring to the classification of stylistics, this study selected four text types: (1) narrative, (2) expository, (3) argumentative, and (4) practical writing. This aims to ensure the diversity of generated text types to cover different reading comprehension situations

and provide sufficient corpus support for subsequent item generation.

3.3 Skill Selection

The effectiveness of automated item generation depends on the generation of items based on specific language skills. Referring to the descriptive framework of the Common European Framework of Reference for Languages (CEFR), this study extracted nine core reading skills, covering key abilities in reading comprehension. The selected skills not only help generate items with theoretical value but also have broad application value in language testing, learning, and teaching practices.

Table 1: *Generated Reading Skills*

No.	Skill Attribute	Definition
1	Infer meaning and use of unfamiliar vocabulary	Able to understand unfamiliar vocabulary based on context and word formation
2	Understand explicitly stated information	Able to locate and understand single explicit information
3	Paraphrase information	Able to restate or interpret text content
4	Analyze multiple details	Able to analyze and summarize information from multiple details
5	Summarize main idea	Able to summarize the main idea of a paragraph or passage
6	Understand discourse cohesion	Able to understand various cohesive and referential relationships
7	Make inferences from content	Able to make inferences about non-explicit information based on content
8	Understand author's purpose, attitude, tone, and emotion	Able to determine the author's purpose, attitude, tone, and emotion based on the entire passage
9	Evaluate author's writing style	Able to evaluate the author's style based on linguistic features and rhetorical devices

3.4 Item Generation

After selecting reading materials and reading skills, the core of the research was the automated generation of items. Referring to the proficiency standards of the Common European Framework of Reference for Languages (CEFR), this study covered six language proficiency levels: A1, A2, B1, B2, C1, and C2 during item generation. For each level, one reading passage was generated for each of the four text

types, resulting in a total of 24 reading passages. For each passage, five multiple-choice questions were generated, totaling 120 items. To ensure that the generated items covered all nine reading skills more evenly, this study adopted a specific research design to ensure a reasonable and representative distribution of items across skill dimensions.

Table 2: *Item Generation Matrix*

Skills and Levels	Corresponding Item Range
Levels A1 and A2 used skills 1–5	Items 1–40
Levels B1 and B2 used skills 3–7	Items 41–80
Levels C1 and C2 used skills 5–9	Items 81–120

3.5 Expert Evaluation

This study used an expert evaluation method to systematically measure the ability of large language models to generate items. A total of five experts with rich experience in language testing participated in the evaluation. They all had professional evaluation literacy and could reliably judge the quality of the generated items. Experts rated each of the 24 generated reading test passages on nine dimensions.

To operationalize the evaluation of generated items, we developed a nine-dimension rating framework. This framework was informed by two complementary sources. First, we drew upon established principles of language testing and multiple-choice item construction, particularly regarding stem clarity, distractor plausibility, single correctness, and alignment with target proficiency levels (Alderson et al., 1995; Bachman & Palmer, 1996; Downing, 2006; Haladyna & Rodriguez, 2013). Second, we incorporated criteria commonly used to assess the quality of AI-generated textual content, including fluency, factual accuracy, relevance to prompts, and safety regarding biased or harmful content (Chang, 2024; Lee et al., 2024; Lin & Chen, 2024). The nine specific dimensions were then synthesized to cover three overarching aspects of item quality: (a) content fidelity (dimensions 1, 7, 8, 9), (b) difficulty and design appropriateness (dimensions 2, 3, 4, 5), and (c) ecological validity relative to real-world examinations (dimension 6).

The nine rating dimensions include:

1. The topic is consistent with the prompt requirements;
2. The text difficulty corresponds to the CEFR reading difficulty level;
3. The question stem is reasonably phrased;
4. There is only one correct answer;
5. The distractors are reasonably set;
6. The item type is similar to that of authoritative exams;
7. The text language is natural and fluent;
8. The text contains no biased or toxic content;
9. The text content is consistent with facts and has no obvious errors.

These nine dimensions are not only important indicators for evaluating the quality of reading texts but also fully reflect the ability of large language models in reading item generation. Regarding specific reading skills, the same five experts were invited to rate all 120 multiple-choice questions. Ratings used a five-point Likert scale, where 1 meant "strongly disagree" and 5 meant "strongly agree."

3.6 Data Analysis

This study used R language for statistical analysis of the expert evaluation data. First, the ratings of the five raters for each item were averaged to reflect the overall performance of each large language model in item generation ability. To further measure the generalization ability of the language models across different proficiency levels and topic types, this study calculated the standard deviation of the ratings for each model. This statistic can effectively reflect the stability and generalizability of the models' generated items. In addition, to compare the differences in automated item generation ability among different language models, this study used the Kruskal-Wallis test for non-parametric comparative analysis of model ratings, thus providing empirical evidence for the selection and application of language models.

IV. Research Results

4.1 Comparison of Large Language Models' Text Content Generation Ability

In terms of content dimension ratings, the performance of the three language models was generally similar (see Table 3). On dimensions such as "topic consistent with prompt requirements," "question stem reasonably phrased," "only one correct answer," "distractors reasonably set," "similar to authoritative exams," "text language natural and fluent," "text contains no biased or toxic content," and "text content consistent with facts, no obvious errors," the average scores of the three models ranged from 3.97 to 5.00, showing a high degree of consistency. The small standard deviations (SD = 0.00–1.12) indicate that the automated item generation performance of the different large language models was relatively stable across levels from A1 to C2.

Table 3: *Comparative Analysis of Large Language Models' Text-Level Generation Ability*

Dimension	Deepseek	GPT-4	ERNIE Bot
Topic consistent with prompt requirements	4.99 (0.09)	4.98 (0.13)	4.89 (0.61)
Text difficulty corresponds to CEFR reading difficulty level	4.30 (1.03)	3.93 (1.03)	4.16 (1.09)
Question stem reasonably phrased	4.36 (0.83)	4.22 (0.89)	4.23 (0.93)
Only one correct answer	4.89 (0.50)	4.72 (0.74)	4.78 (0.74)
Distractors reasonably set	4.26 (0.85)	4.05 (1.12)	3.97 (1.11)
Similar to authoritative exams	3.87 (0.92)	3.84 (0.96)	3.73 (1.06)
Text language natural and fluent	4.73 (0.53)	4.68 (0.58)	4.81 (0.42)
Text contains no biased or toxic content	5.00 (0.00)	4.99 (0.09)	5.00 (0.00)
Text content consistent with facts, no obvious errors	5.00 (0.00)	4.98 (0.13)	4.98 (0.18)

Note. Values are means (M), with standard deviations (SD) in parentheses.

Notably, on the dimension "text difficulty corresponds to CEFR reading difficulty level," the scores of the models differed somewhat. Deepseek scored the

highest (M = 4.30, SD = 1.03), followed by ERNIE Bot (M = 4.16, SD = 1.09), and GPT-4 scored the lowest (M = 3.93, SD = 1.03), indicating that this dimension is more sensitive to differences between models.

To further analyze the differences in automated item generation ability among the different models, this study conducted Kruskal-Wallis tests on the scores of each dimension. The results showed that, except for "text difficulty corresponds to CEFR," the differences in the other eight dimensions were not significant, indicating that the performance of the different models on these dimensions was relatively similar. However, on the "text difficulty corresponds to CEFR" dimension, the difference among models reached a significant level ($\chi^2 = 15.14$, $p < 0.001$). Post-hoc comparisons showed that Deepseek's score (M = 4.30, SD = 1.03) was significantly higher than GPT-4's (M = 3.93, SD = 1.03, $p < 0.001$), and GPT-4's score was significantly lower than ERNIE Bot's (M = 4.16, SD = 1.09, $p < 0.05$); the difference between Deepseek and ERNIE Bot was not significant ($p > 0.05$).

Table 4: *Kruskal-Wallis Test Comparison Results*

No.	Dimension	N	χ^2	p-value
1	Topic consistent with prompt requirements	362	2.10	0.350
2	Text difficulty corresponds to CEFR reading difficulty level	362	15.14	0.001**
3	Question stem reasonably phrased	362	1.79	0.408
4	Only one correct answer	362	5.15	0.076
5	Distractors reasonably set	362	3.06	0.217
6	Similar to authoritative exams	362	0.72	0.696
7	Text language natural and fluent	362	2.40	0.301
8	Text contains no biased or toxic content	362	2.00	0.368
9	Text content consistent with facts, no obvious errors	362	1.99	0.370

Note. **p < 0.001.

Overall, the Kruskal-Wallis effect size analysis indicated that the differences among the three models on various dimensions of item quality were not significant, especially on dimensions such as reasonableness of question stems, uniqueness of correct answer, naturalness of text language, and bias/factual errors. On the reading difficulty matching dimension, although there were small differences, they did not

reach a medium effect level.

4.2 Comparison of Large Language Models' Ability to Generate Items Based on Reading Skills

In terms of reading skill dimensions, the overall performance of the three language models was relatively close (Table 5). On dimensions such as "understand author's purpose, attitude, tone, and emotion," "summarize main idea," "understand explicitly stated information," and "understand discourse cohesion," the average scores of the three models were all between 4.3 and 4.8, showing little difference, indicating that the three models had relatively consistent generation quality on these reading comprehension skills.

Table 5: *Scores of Different Language Models on Different Skills*

Skill Name	Deepseek	GPT-4	ERNIE Bot
Understand author's purpose, attitude, tone, and emotion	4.70 (0.40)	4.55 (0.21)	4.60 (0.30)
Make inferences from content	4.44 (0.39)	4.21 (0.56)	3.90 (0.59)
Paraphrase information	4.40 (0.60)	4.30 (0.62)	4.22 (0.40)
Analyze multiple details	3.95 (0.51)	4.03 (0.59)	4.26 (0.57)
Summarize main idea	4.75 (0.36)	4.72 (0.22)	4.70 (0.33)
Infer meaning and use of unfamiliar vocabulary	4.72 (0.37)	4.95 (0.09)	4.62 (0.47)
Understand explicitly stated information	4.75 (0.32)	4.85 (0.21)	4.72 (0.26)
Understand discourse cohesion	4.45 (0.34)	4.39 (0.33)	4.33 (0.44)
Evaluate author's writing style	4.83 (0.17)	4.62 (0.27)	4.47 (0.40)

Note. Values are means (M), with standard deviations (SD) in parentheses.

There were some differences among models on certain skill dimensions. For example, on the dimension "make inferences from content," Deepseek had the highest

average score (M = 4.44, SD = 0.39), followed by GPT-4 (M = 4.21, SD = 0.56), and ERNIE Bot the lowest (M = 3.90, SD = 0.59), showing a gradient difference. On the dimensions "paraphrase information" and "analyze multiple details," the scores of the three models were relatively close, but ERNIE Bot was slightly higher than the other models on "analyze multiple details" (M = 4.26, SD = 0.57).

In addition, on the dimension "infer meaning and use of unfamiliar vocabulary," GPT-4 scored the highest (M = 4.95, SD = 0.09), significantly higher than Deepseek (M = 4.72, SD = 0.37) and ERNIE Bot (M = 4.62, SD = 0.47). On the dimension "evaluate author's writing style," Deepseek scored the highest (M = 4.83, SD = 0.17), followed by GPT-4 (M = 4.62, SD = 0.27), and ERNIE Bot the lowest (M = 4.47, SD = 0.40).

To further compare the generation ability of different language models on different skills, the Kruskal-Wallis test was also used for comparative analysis. The results showed that there were significant differences among the three models on the skill "make inferences from content" (see Table 6). On this skill dimension, the Kruskal-Wallis test showed significant differences among the three models ($\chi^2 = 7.14$, $p = 0.028$). Further pairwise comparisons revealed: a significant difference between Deepseek and ERNIE Bot (p.adj = 0.022), with Deepseek's average rating higher than ERNIE Bot's; the difference between Deepseek and GPT-4 was not significant (p.adj = 0.295); the difference between GPT-4 and ERNIE Bot was also not significant (p.adj = 0.213). In summary, the significant difference on this skill dimension mainly existed between Deepseek and ERNIE Bot, indicating that Deepseek is slightly better than ERNIE Bot in the ability to "make inferences from content," while performing similarly to GPT-4.

Table 6: *Comparative Analysis of Language Models on Different Reading Skills*

Skill Dimension	n	Kruskal-Wallis Statistic	df	p-value
Understand author's purpose, attitude, tone, and emotion	24	2.38	2	0.304
Make inferences from content	48	7.14	2	0.0282*
Paraphrase information	48	2.16	2	0.339

Analyze multiple details	48	2.66	2	0.264
Summarize main idea	72	1.89	2	0.388
Infer meaning and use of unfamiliar vocabulary	24	3.26	2	0.196
Understand explicitly stated information	24	1.17	2	0.558
Understand discourse cohesion	48	0.494	2	0.781
Evaluate author's writing style	24	4.33	2	0.115

Note. *p < 0.05.

4.3 Comparison of Large Language Models' Ability to Generate Items Based on Reading Levels

Table 7 presents the nine-dimensional comprehensive rating results of the models' generated reading items across six levels. Looking at the trend across levels, DeepSeek maintained high scores at all levels and reached its highest value at level C2 (M = 42.9, SD = 1.52), indicating that it has stronger item generation ability and consistency in high-level reading tasks. ChatGPT achieved its highest score at level A2 (M = 42.3, SD = 2.83), but fluctuated slightly in the intermediate and high levels (B1–B2), reflecting its insufficient control over intermediate-difficulty texts. ERNIE Bot fluctuated more at level B1 (SD = 5.74) but became stable at high levels (C1–C2), performing similarly to DeepSeek.

Table 7: *Comparison of Model Performance in Generating Reading Items by CEFR Level*

Level	ChatGPT	DeepSeek	ERNIE Bot	H	p-value
A1	40.40 ± 3.44	41.35 ± 2.98	40.70 ± 4.51	0.62	0.735
A2	42.30 ± 2.83	40.30 ± 5.03	40.40 ± 4.97	1.53	0.466
B1	39.90 ± 4.77	41.20 ± 2.67	38.55 ± 5.74	0.90	0.638
B2	39.60 ± 4.30	41.20 ± 2.89	40.50 ± 2.96	1.88	0.391
C1	40.15 ± 4.16	41.45 ± 2.65	41.45 ± 1.57	0.66	0.719
C2	40.10 ± 3.97	42.90 ± 1.52	41.75 ± 2.75	10.04	0.007**

Note. Values are means ± standard deviations (M ± SD). Significance tests based on the Kruskal-Wallis H test. **p < 0.01.

From the statistical results, the Kruskal-Wallis H test showed that except for level C2, the differences among the three models at other levels were not significant, indicating that their generation abilities were comparable on low- and intermediate-

level tasks. However, at level C2, the difference was significant ($H = 10.04$, $p = 0.007$). Further Dunn's post-hoc test revealed that DeepSeek and ChatGPT differed significantly at level C2 ($p = 0.002$), while the differences between DeepSeek and ERNIE Bot and between ChatGPT and ERNIE Bot were not significant. This result indicates that DeepSeek is significantly better than ChatGPT in generating high-level reading items, with ERNIE Bot performing between the two.

V. Discussion

This study aimed to explore the differences in automated item generation ability among large language models. The results showed that on the dimension "text difficulty corresponds to CEFR reading difficulty level," Deepseek performed significantly better than GPT-4, and ERNIE Bot also performed better than GPT-4. This indicates that the domestic large language model Deepseek has achieved excellent results in generation ability, even surpassing the international mainstream model GPT-4. This finding differs from previous research (Lin & Chen, 2024) in that it provides a reference for automated item generation from the perspective of model performance comparison.

This advantage may be closely related to Deepseek's training method, which emphasizes the learning of reasoning processes, thereby improving its ability to understand user needs and discourse logical structures. In the comparison of skill dimensions, the research results remained consistent with this trend. For example, on the core reading skill "make inferences from content," Deepseek also performed better than GPT-4 and ERNIE Bot, showing its clear advantage in reasoning tasks. This indicates that domestic large language models have significant potential for generating inference-type items, especially suitable for language testing scenarios that require deep understanding and logical deduction.

Different from previous studies (Sayin & Gierl, 2024; Shin et al., 2025), this study comprehensively evaluated the performance of large language models in

automated item generation at different language proficiency levels (A1 to C2). By analyzing the variance of ratings, the study quantified the performance fluctuations and stability of the models across different levels. The results showed that the standard deviations of the ratings on various dimensions were small, indicating that large language models have good generalization ability and stability in automated item generation across different topics and levels. This finding provides empirical support for the feasibility, generalizability, and reliable application of automated item generation under different difficulty and topic conditions, and also provides a reference for the design of multi-level language tests in the future.

This study also found that on the dimension "similar to authoritative exams," the scores of the three large language models were relatively low, indicating that there is considerable room for improvement in this dimension. Possible reasons include: first, the item type used in this study was only multiple-choice, which did not cover other item types in authoritative exams, which may have limited the ability of the models to generate items highly consistent with real exams; second, it may also reflect that these models still have room for improvement in item generation. Future research could further verify the reliability of the findings of this study by comparing the quality of machine-generated and human-written items, and provide a reference for optimizing the performance of large language models in automated item generation.

In addition, whether the content generated by large language models contains bias or false information is a potential risk that must be focused on in language testing applications (Chang, 2024). The results of this study showed that no such problems were found in the 120 passages generated in this study, indicating that large language models can generate relatively safe and reliable test content in the preliminary application stage. However, this study relied only on expert evaluation to determine content bias. In practical applications, further quantitative methods, such as differential item functioning (DIF) analysis or text bias detection algorithms, can be used to systematically verify the content generated by the models to ensure its fairness and accuracy.

It is worth noting that this study mainly focused on the ability of large language models at the text generation level, but their potential applications go far beyond that. Future research could further explore the ability of models to intelligently adapt existing texts, such as adjusting text difficulty, optimizing vocabulary and syntactic structures, and enhancing logical coherence. Such functions can not only provide more flexible item design solutions for language testing but can also be widely used in language learning and teaching practices, supporting the development of personalized learning paths and adaptive assessments, thereby expanding the application prospects of large language models in the field of educational technology.

VI. Conclusion

This study systematically evaluated the performance of large language models in automated item generation through expert evaluation. The results showed that Deepseek, GPT-4, and ERNIE Bot generally had little difference in text generation ability, but Deepseek performed better in generating specific text difficulties. In terms of skill dimensions, the three large language models performed similarly overall, but Deepseek performed better in generating inference-type items. These results indicate that this study can provide important references for model selection, item construct design, and text generation in large language model-based automated item generation, while also providing theoretical and practical foundations for promoting the practical application of large language models in language testing.

This study also has some limitations. First, the study focused on the generation ability of multiple-choice questions, lacking diversity in item types. Future research could expand to other item types, such as matching, fill-in-the-blank, or short-answer questions, to comprehensively evaluate the generation ability of large language models on different item types. Second, this study invited only five experts to participate in the evaluation, a relatively small sample size. Future research could increase the number of experts or combine cross-institutional reviews to enhance the generalizability and reliability of the findings. Finally, this study mainly relied on expert

judgment to evaluate generation ability. Future research could combine actual test data and use psychometric models (e.g., IRT models) or other quantitative methods to more systematically evaluate the validity, discrimination, and fairness of items generated by large language models, thus providing a more solid empirical foundation for the application of automated item generation in language testing.

References

Ahmed, A., Jamil, E., Abubakar, M., Batool, A., Masoom Akhtar, M., Iqbal Nasiri, M., & Ullah, M. (2024). Harnessing the power of ChatGPT to develop effective MCQ-based clinical pharmacy exams. *Journal of Research on Technology in Education*, 1-11. https://doi.org/10.1080/15391523.2024.2425435

Alderson, J. C., Clapham, C., & Wall, D. (1995). *Language test construction and evaluation*. Cambridge University Press.

Aryadoust, V., Zakaria, A., & Jia, Y. (2024). Investigating the affordances of OpenAI's large language model in developing listening assessments. *Computers and Education: Artificial Intelligence*, *6*, 100204. https://doi.org/https://doi.org/10.1016/j.caeai.2024.100204

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice*. Oxford University Press.

Chang, H.-H. (2024). Harnessing AI for Educational Measurement: Standards and Emerging Frontiers. *Journal of Educational and Behavioral Statistics*, 10769986241264033. https://doi.org/10.3102/10769986241264033

Downing, S. M. (2006). Selected-response item formats in test development. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 287-301). Lawrence Erlbaum.

Haladyna, T. M., & Rodriguez, M. C. (2013). *Developing and validating test items*. Routledge.

Lin, Z., & Chen, H. (2024). Investigating the capability of ChatGPT for generating multiple-choice reading comprehension items. *System*, 103344. https://doi.org/10.1016/j.system.2024.103344

Liu, Y., Bhandari, S., & Pardos, Z. A. (2025). Leveraging LLM respondents for item evaluation: A psychometric analysis. *British Journal of Educational Technology*, *n/a*(n/a). https://doi.org/https://doi.org/10.1111/bjet.13570

Sayin, A., & Gierl, M. (2024). Using OpenAI GPT to Generate Reading Comprehension Items. *Educational Measurement: Issues and Practice*. https://doi.org/10.1111/emip.12590

Shin, D., Lee, J. H., & Kim, K. (2025). An Exploratory Study on Two Automated Item Generators for Generating L2 Reading Test Items. *RELC Journal*. https://doi.org/10.1177/00336882251326284

White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., Gilbert, H., Elnashar, A., Spencer-Smith, J., & Schmidt, D. C. (2023). A prompt pattern catalog to enhance prompt engineering with ChatGPT. *arXiv preprint arXiv:2302.11382*.

Yang., A. C., Chen, I. Y., Flanagan, B., & Ogata, H. (2021). Automatic generation of cloze items for repeated testing to improve reading comprehension *Educational Technology & Society*, *24*(3), 147-158.

Zhu, J., Liu, D., & Chen, S. (2024). Multiple-choice question generation and difficulty calculations based on semantic similarity. *Neural Computing and Applications*. https://doi.org/10.1007/s00521-024-10671-8