Proceedings of the 90th Annual International Meeting of the Psychometric Society, Minneapolis, United States (2025).

Automated Item Generation for Reading Assessments

Ummugul Bezirhan [*†], Erin Wry [†], Sebahat Goren [‡], and Matthias von Davier [†]

†Boston College, Chestnut Hill, 02467, MA, USA
‡Kütahya Dumlupınar University, Kütahya, Türkiye
*Corresponding author. Email: bezirhan@bc.edu

Abstract

This study explores the application of automated item generation (AIG) using large language models (LLM) within the PIRLS (Progress in International Reading Literacy Study) reading assessment framework. By operationalizing PIRLS item-writing guidelines via GPT-based prompting strategies, the study demonstrates a scalable approach to developing reading-comprehension items. The item generation process involved two main steps: first, generating passage-specific items using structured prompts; second, filtering items using similarity metrics. After this automated filtering, experts reviewed the remaining items to ensure content quality and alignment with PIRLS standards and curated the set. As a final step, the items were evaluated psychometrically against human-authored items. The results suggest that LLM-based AIG can be a valuable tool to support item development in large-scale international reading assessments.

Keywords: automated item generation, large language models, ILSAs

1. Introduction

With the widespread adoption of computer-based testing, the demand for high-quality items has grown substantially. Traditional item development methods, however, often fall short of meeting this demand. These methods are among the most costly, time-consuming, and labor-intensive aspects of assessment development because they rely heavily on subject-matter experts (SMEs), who are responsible for drafting each item and conducting multiple rounds of review and revision to ensure quality and alignment with content standards (Haladyna & Rodriguez, 2013).

AIG offers new possibilities for assessment development by integrating principles from cognitive science and measurement theory with computational techniques (e.g., Gierl & Haladyna, 2012). Early AIG research relied on template-based frameworks where SMEs defined item models and variable elements to generate a large number of items. Although effective, these approaches were constrained by limited linguistic variability and continued dependence on SMEs. Subsequent work introduced natural language processing (NLP)- based methods, followed by neural network approaches that learned item patterns directly from large corpora and enabled greater linguistic diversity in generated items (e.g., Flor & Riordan, 2018; Shin, 2021; von Davier, 2018).

The rapid progress in NLP following the introduction of transformer architectures (Vaswani et al., 2017) has prompted growing interest in the feasibility of using LLMs for assessment development. Accordingly, several researchers have proposed incorporating LLMs into AIG pipelines (e.g., Attali et al., 2022; Bezirhan & von Davier, 2023; von Davier, 2019), especially in assessment domains where traditional template-based or rule-driven AIG methods have proven difficult to implement at scale.

Reviews of AI-based AIG consistently indicate that prompt engineering has become the dominant mechanism for controlling LLM output (Kurdi et al., 2020; Song et al., 2025; Tan et al., 2024). In contrast to earlier template- or rule-based approaches, prompt-driven AIG relies on natural-language

specifications to encode item-writing rules, cognitive targets, and design constraints. However, these reviews also note that many empirical studies rely on poorly specified prompts and evaluate generated items primarily through expert judgment, often without explicit reference to assessment frameworks. Thus, the resulting items may exhibit construct underrepresentation and misalignment with intended comprehension processes. Studies that embed framework-specific constraints directly into prompts and combine generation with human-in-the-loop review report more consistent item or content quality (e.g., Attali et al., 2022; Bezirhan & von Davier, 2023; Sayın & Gierl, 2024; Wen & Chu, 2025). These findings suggest that prompt design functions as the primary mechanism through which assessment theory is translated into LLM behavior. Accordingly, prompts that merely request "good" or "high-quality" questions tend to yield heterogeneous output. In contrast, prompts that specify roles, blueprint constraints, output schemas, and explicit alignment checks produce more stable and usable item sets. In this sense, reading comprehension items are particularly demanding because validity depends on text dependence, ordered progression, balanced difficulty, and cognitive process targeting.

PIRLS provides a rigorous framework for evaluating reading comprehension in Grade 4 students. PIRLS focuses on two purposes for reading: for literary experience (literary) and to acquire and use information (informational), with items targeting four broad-based comprehension processes: focus on and retrieve explicitly stated information, make straightforward inferences, interpret and integrate ideas, and evaluate and critique content and textual elements (Wry & Mullis, 2021). Items must be anchored in the passage, follow its progression, and remain independent of one another. These characteristics make PIRLS an ideal context for assessing whether LLM-generated items meet the item-quality standards required for international large-scale assessments. While recent work has shown that LLMs can generate reading passages comparable to human-authored PIRLS texts (Bezirhan & von Davier, 2023), limited research has examined item-level generation aligned with key comprehension processes and evaluated using student response data. The present study addresses this gap by developing an AIG workflow combining structured prompt design, algorithmic filtering, expert review, and psychometric field testing.

Beyond evaluating item quality, a central aim of this study is to examine whether LLM-generated items can support the construction of alternate assessment forms – items that are conceptually aligned with assessment targets while remaining distinct from existing questions. Such alternate forms are essential for broader construct coverage, test security, and, in the longer term, may serve as a foundation for adaptive implementations of PIRLS. The goal is not to replace expert item writers but to examine the extent to which AI can support and scale item development.

2. Methodology

Two released PIRLS 2021 reading passages, one informational (Passage I) and one literary (Passage II), were selected as source material. The item generation process began with prompt development informed by PIRLS item-writing guidelines that are used in the actual item development process. Initially, a general prompt was used to generate a broad set of questions for each passage. Although this approach yielded volume, expert feedback indicated that it struggled to capture the nuances of the passages necessary to address all four comprehension processes specified in the PIRLS reading framework. In response to this feedback, the procedure was refined into a process-specific prompting strategy. For each of the four comprehension processes targeted in PIRLS, an automated, process-specific prompt was used, in which the descriptions from the PIRLS framework for the corresponding process were programmatically embedded. These prompt constraints guided GPT-4o to generate items that were text-dependent, cognitively targeted, and consistent with PIRLS item-writing conventions. Prompts were generated via an API call, with 5 replications at a temperature setting of 1 to maximize linguistic diversity, yielding approximately 150 items per passage.

The resulting item pool contained both exact duplicates and semantically overlapping paraphrases.

As a first filtering step, exact duplicate items were removed using cosine similarity. Because the objective was to construct an alternate form rather than replicate the original assessment, a second filtering step focused on identifying AI-generated items that were overly similar to the human-authored PIRLS questions for the same passages. Items that closely mirrored existing PIRLS items simultaneously in content, structure, and comprehension process were excluded. Importantly, items that shared similar content with PIRLS questions but differed in response format, structural features or targeted comprehension process were retained, reflecting common practices in test development.

To support this review, a multi-metric similarity evaluation was conducted to compare the remaining AI-generated items with the original PIRLS items. BLEU scores assessed n-gram precision, ROUGE-L evaluated longest common subsequence alignment, and METEOR accounted for synonym and paraphrase matching. In addition, cosine similarity based on Sentence-BERT embeddings yielded semantic similarity independent of the surface structure. A weighted ranking system was applied, assigning 40% to cosine similarity, 30% to ROUGE-L, and 30% to METEOR, to identify items conceptually aligned with PIRLS targets while avoiding near-duplicate items of the original items. BLEU scores were examined for diagnostic purposes but were not weighted heavily due to their sensitivity to exact lexical overlap.

The filtering process reduced the pool to 45 candidates for Passage I and 49 for Passage II. A reading expert reviewed the remaining items to evaluate theoretical alignment, clarity, distractor plausibility, and developmental appropriateness. From these pools, 14 items for Passage I and 17 items for Passage II were selected to create alternate forms comparable in breadth to the original PIRLS blocks. All items were translated from English to Turkish by a native speaker and validated by an expert Turkish item writer to ensure clarity, cultural appropriateness, and comparability to the international source version.

The AI-generated alternative items were administered to Grade 4 students in Türkiye, a PIRLS 2021 participating country. For Passage I, responses were collected from approximately 115 students per item, and for Passage II from approximately 200 students per item. Item difficulty, operationalized as the proportion of correct responses, and discrimination indices, calculated through item-total correlations, were compared with official PIRLS 2021 statistics to evaluate functional equivalence.

3. Results

3.1 Item Selection Process

Using two released PIRLS 2021 passages, one informational and one literary, over 150 AI-generated questions per passage were produced. After removing duplicates, a filtering strategy was applied using empirically determined thresholds for BLEU, ROUGE, METEOR, and cosine similarity to optimize quality and recall.

Figure 2 shows the distribution of similarity scores across four metrics. BLEU measures exact n-gram matches and shows near-zero values for most items, indicating AI-generated items rarely replicate exact phrasing from PIRLS questions. This is expected, as multiple valid phrasing options exist. METEOR, which accounts for synonyms and paraphrasing, yielded slightly higher values (0-0.2), though the right-skewed distribution indicates semantic alignment remained a challenge. ROUGE-L, measuring longest common subsequences, showed moderate values, reflecting varied sentence structures. Cosine Similarity exhibited a bimodal distribution with a major peak at approximately 0.6, suggesting that although AI items used different phrasing, they often captured similar semantic meaning.

Figure 3 illustrates the effect of the filtering strategy on the distributions of metrics. BLEU showed a modest rightward shift, reflecting the removal of items with negligible n-gram overlap. ROUGE-L transformation was particularly striking, with the post-filtering distribution extending into higher values (0.2-0.8), indicating strong prioritization of structural similarity. METEOR scores shifted significantly rightward (0.2-0.4 range), demonstrating success in retaining semantically aligned

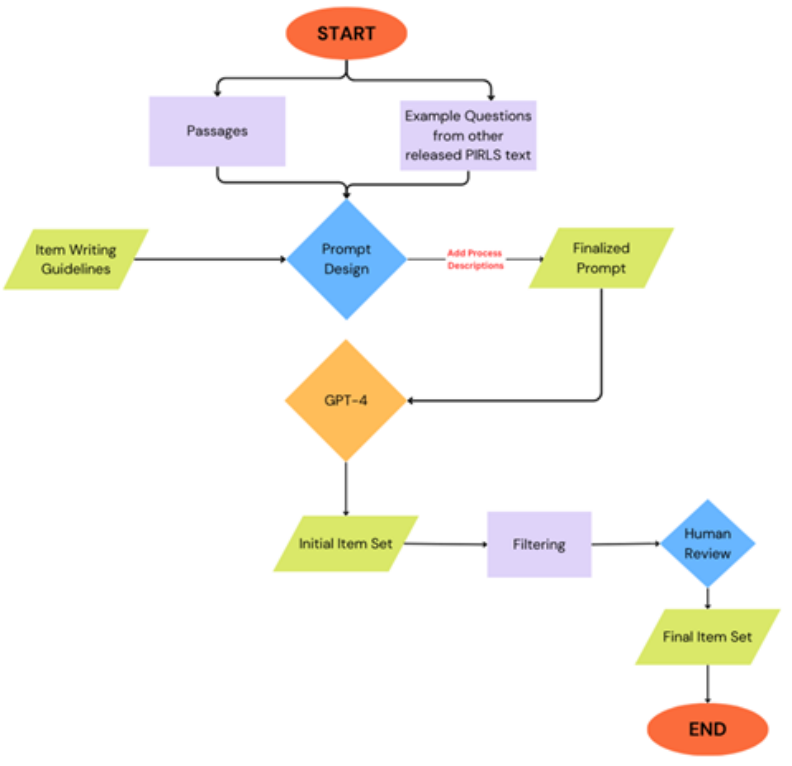

Figure 1. Pipeline used for AIG

items. Cosine Similarity maintained its bimodal character but concentrated predominantly in the higher range (above 0.5). These distributions validate the threshold-based approach and weighted ranking system (40% Cosine, 30% ROUGE-L, 30% METEOR). The consistent rightward shifting demonstrates that the methodology effectively aligned the candidate pool with PIRLS reference questions while maintaining linguistic diversity.

Following quantitative filtering, a PIRLS content expert reviewed the respective item sets to ensure that the alternate forms were distinct from the original PIRLS blocks corresponding to the same reading passages. This qualitative validation addressed limitations of computational metrics: identifying nuanced similarity beyond lexical and semantic measures, ensuring coverage of all four comprehension processes assessed in PIRLS, and validating item authenticity. The expert selected a set of items for each passage that maintained structural equivalence with the original PIRLS blocks - 14 items for Passage I (original: 15 items) and 17 for Passage II (original: 17 items). This human-in-the-loop validation bridges computational generation and practical assessment development, ensuring alternate forms aligned with PIRLS objectives while appropriately differentiating from source items.

3.2 Psychometric Comparison

Figure 4 presents the psychometric comparison for Passage I. AI-generated items (14 items) showed slightly lower difficulty (M = 68.5%, SD = 9.1) compared to PIRLS items (15 items; M = 71.3%, SD = 14.4). The narrower standard deviation for AI items suggests more consistent difficulty targeting. Discrimination values for AI items (M = 0.38, SD = 0.12) were moderately lower than PIRLS items (M = 0.47, SD = 0.06). Notably, all AI items exceeded the marginally acceptable discrimination threshold of 0.2 (Allen & Yen, 2001; Ebel & Frisbie, 1991), with the range extending to 0.68. Alternatively, the

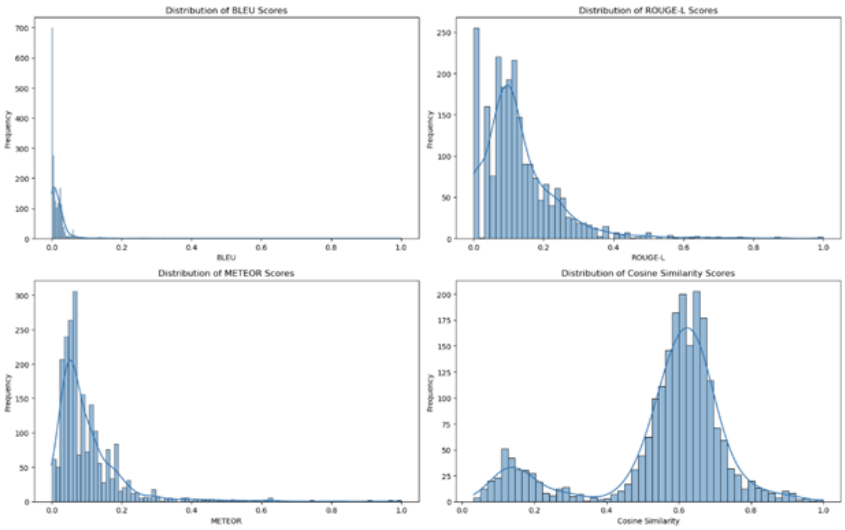

Figure 2. Distribution of BLEU, ROUGE-L, METEOR, and Cosine Similarity scores before filtering (Passage I). Passage II exhibited comparable metric distributions.

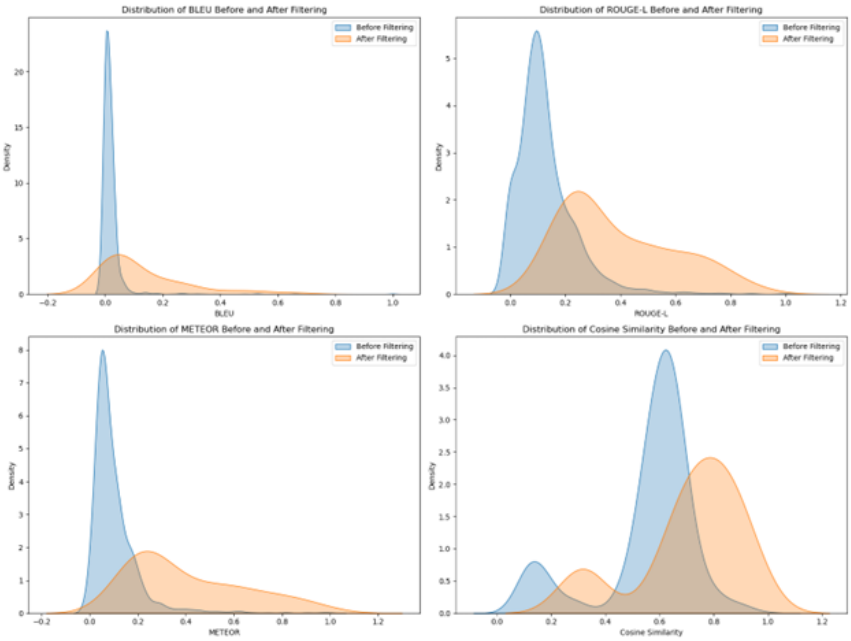

Figure 3. Effect of filtering on similarity metric distributions (Passage I). Similar transformations were observed for Passage II.

Passage II results, shown in Figure 5, indicate that the AI-generated items were more difficult (M = 59.5%, SD = 15.2) than the PIRLS items (M = 67.0%, SD = 10.7), with notably greater variability in difficulty. For Passage II, AI items also demonstrated greater discrimination (M = 0.48, SD = 0.11) than PIRLS items (M = 0.40, SD = 0.07), suggesting greater differentiation between high- and low-performing students. Both item sets fell within acceptable psychometric ranges. These results indicate that AI-generated items function effectively in measuring reading comprehension across both passages. While exhibiting passage-specific variation in psychometric properties, all items demonstrated acceptable difficulty and discrimination parameters comparable to expert-developed PIRLS items.

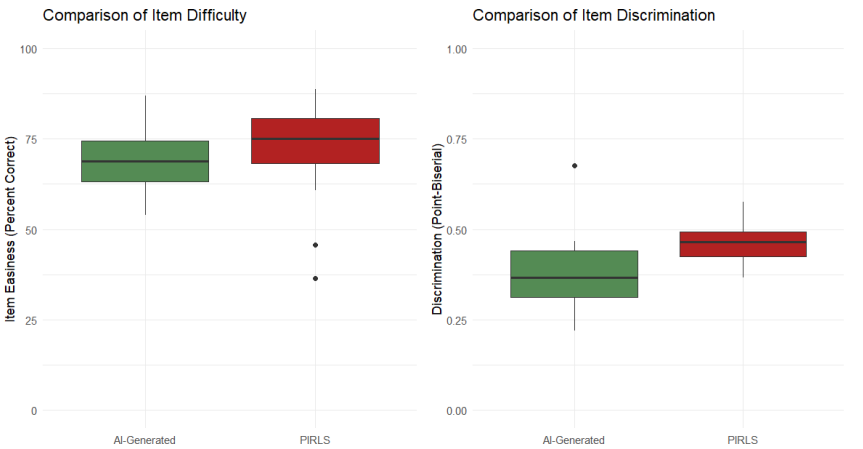

Figure 4. Difficulty and Discrimination Comparison for Passage I.

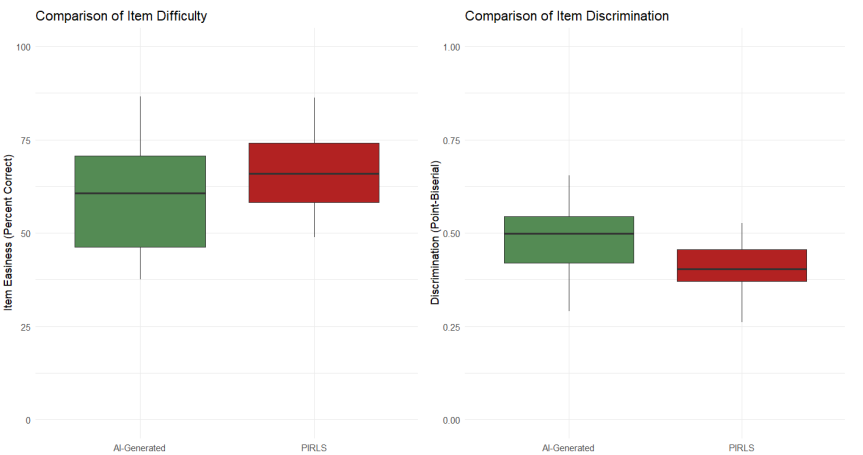

Figure 5. Difficulty and Discrimination Comparison for Passage II.

4. Discussion

The results suggest that LLM-based item generation, when guided by structured prompt architectures and combined with systematic similarity filtering and expert validation, can yield reading

comprehension items that function comparably to human-generated PIRLS questions. The workflow presented here significantly reduces the initial burden on SMEs by shifting their role from item writing to evaluative decision-making. Instead of writing items from scratch, experts review and refine algorithm-produced candidates, resulting in notable resource efficiency while retaining control over construct and framework alignment.

It is important to note that the high-quality outcomes described depended heavily on carefully engineered prompts and expert review. Without explicit framework encoding, LLMs tended to generate generic, text-adjacent questions lacking theoretical specificity. Furthermore, automated similarity metrics, although effective for initial filtering, require further refinement to reliably detect distractor weakness, comprehension process misalignment, and subtle construct drift. In particular, generating effective distractors for passage-based items presents a distinct challenge: in PIRLS, response options are grounded directly in passage content, requiring distractors that reflect plausible but passage-specific misinterpretations. LLMs may default to more generic options that while functionally acceptable lack the nuance needed to produce consistent discrimination across items. Additionally, the weighting scheme applied in the filtering stage was informed by the complementary properties of each metric rather than a formal sensitivity analysis; future work should examine how alternative weighting configurations affect the composition and quality of the final item set. These findings reinforce the view that AI is best positioned as a collaborator rather than a replacement in assessment development.

The psychometric evidence is particularly encouraging given the methodological constraints of this study. Importantly, the prefiltering and expert review phase explicitly excluded items that were excessively similar to the original, high-quality PIRLS questions, thereby ensuring that alternate forms were created rather than replicas of existing assessments. This requirement made the validation task considerably more challenging, as additional items needed to be identified that were not similar to existing ones; AI-generated items needed to demonstrate psychometric soundness while remaining relatively distinct from reference items. Despite this constraint, results from both passages demonstrated that AI-generated items functioned within acceptable psychometric ranges. For Passage I, AI items showed slightly lower discrimination than PIRLS items, while for Passage II, AI items achieved higher discrimination. The variation across passages may indicate that item quality depends on the interaction between passage characteristics and prompt design. One tentative explanation is that literary passages, with their richer interpretive content and inferential layers, may align particularly well with the kinds of comprehension distinctions LLMs are trained to recognize, enabling more discriminating items. However, the mechanisms underlying this passage-level variation remain an open question that warrants systematic investigation across a broader and more diverse set of passages. Future research should focus on extending this workflow by using larger, more diverse validation samples, incorporating multi-rater expert review to assess inter-rater consistency, and examining the generalizability of the approach across additional passages, grade levels, and languages.

5. Conclusion

This study provides initial evidence that LLMs can contribute productively to item development for international reading assessments. By embedding the PIRLS framework into a multi-step generation and validation pipeline, we demonstrated that AI-produced items can serve as viable alternates to existing questions and perform within acceptable psychometric ranges when administered to students. These results point toward a practical path for integrating generative AI into assessment programs, opening opportunities for increased item diversity and scalability while maintaining alignment with established constructs.

References

Allen, M. J., & Yen, W. M. (2001). *Introduction to measurement theory*. Waveland Press.

Attali, Y., Runge, A., LaFlair, G. T., Yancey, K., Goodwin, S., Park, Y., & von Davier, A. A. (2022). The interactive reading task: Transformer-based automatic item generation. *Frontiers in Artificial Intelligence*, *5*, 903077. https://doi.org/10.3389/frai.2022.903077

Bezirhan, U., & von Davier, M. (2023). Automated reading passage generation with openai's large language model. *Computers and Education: Artificial Intelligence*, *5*, 100161. https://doi.org/10.1016/j.caeai.2023.100161

Ebel, R. L., & Frisbie, D. A. (1991). *Essentials of educational measurement* (5th ed.). Prentice-Hall.

Flor, M., & Riordan, B. (2018). A semantic role-based approach to open-domain automatic question generation. *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, 254–263.

Gierl, M. J., & Haladyna, T. M. (2012). *Automatic item generation: Theory and practice*. Routledge. https://doi.org/10.4324/9780203803912

Haladyna, T. M., & Rodriguez, M. C. (2013). *Developing and validating test items*. Routledge. https://doi.org/10.4324/9780203850381

Kurdi, G., Leo, J., Parsia, B., Sattler, U., & Al-Emari, S. (2020). A systematic review of automatic question generation for educational purposes. *International Journal of Artificial Intelligence in Education*, *30*(1), 121–204. https://doi.org/10.1007/s40593-019-00186-y

Sayın, A., & Gierl, M. (2024). Using openai gpt to generate reading comprehension items. *Educational Measurement: Issues and Practice*, *43*(1), 5–18. https://doi.org/10.1111/emip.12590

Shin, E. (2021). *Automated item generation by combining the non-template and template-based approaches to generate reading inference test items* [Doctoral dissertation]. University of Alberta. https://doi.org/10.7939/r3-75wr-hc80

Song, Y., Du, J., & Zheng, Q. (2025). Automatic item generation for educational assessments: A systematic literature review. *Interactive Learning Environments*, *33*(9), 5386–5405. https://doi.org/10.1080/10494820.2025.2482588

Tan, B., Armoush, N., Mazzullo, E., Bulut, O., & Gierl, M. (2024). A review of automatic item generation techniques leveraging large language models. *International Journal of Assessment Tools in Education*, *12*(2), 317–340. https://doi.org/10.21449/ijate.1602294

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *30*.

von Davier, M. (2018). Automated item generation with recurrent neural networks. *Psychometrika*, *83*(4), 847–857. https://doi.org/10.1007/s11336-018-9608-y

von Davier, M. (2019). Training optimus prime, md: Generating medical certification items by fine-tuning openai's gpt-2 transformer model. *arXiv preprint arXiv:1908.08594*. https://arxiv.org/abs/1908.08594

Wen, Z., & Chu, S. K. W. (2025). Using generative ai for reading question creation based on pirls 2011 framework. *Cogent Education*, *12*(1), 2458653. https://doi.org/10.1080/2331186X.2025.2458653

Wry, E., & Mullis, I. V. S. (2021). Developing the PIRLS 2021 achievement instruments. In M. von Davier, I. V. S. Mullis, B. Fishbein, & P. Foy (Eds.), *Methods and procedures: PIRLS 2021 technical report* (pp. 1–24). TIMSS & PIRLS International Study Center.

Appendix

```
Generate a comprehensive set of questions based on the provided text.  These
questions should be crafted specifically to evaluate the reading comprehension
skills of 4th-grade students regarding the content and various elements within
the text.  Questions should be asked in the same sequential order as the text.
Students should have to read the text in order to answer the question.

Guidelines for Generating Questions:
<Insert process specific guidelines here>

Examples:
<Insert process specific example characteristics here>

Item Types & Specific Guidelines:

Multiple-Choice (1 point):
Stem:  All stems should be phrased as a direct, unambiguous question, avoiding
extraneous information.
Options:  Four choices, labeled A-D.
Key:  The correct answer.
Distractors:  Incorrect options.  They should be plausible and closely
related to the text's ideas or events and crafted based on likely student
errors or misunderstandings about the text.  Ensure options are independent,
grammatically consistent with the stem, and of parallel length.  Do not use
negative stems, but, if necessary, emphasize the negative word.

Constructed Response (1-3 points):
Communicate expectations clearly.  Use definitive terms like ''explain'' or
''describe'' and specify the extent of the expected answer.  The response space
should hint at the desired response length.
Scoring:  3 points for full, correct answers; 1-2 points for partial
correctness; 0 for incorrect answers.
```

Figure A1. Prompt Structure