Input Optimization for Automated Scoring in Reading Assessment

Ji Yoon Jung Ummugul Bezirhan Matthias von Davier
TIMSS & PIRLS International Study Center at Boston College
{jiyoon.jung, bezirhan, vondavim}@bc.edu

Abstract

This study examines input optimization for enhanced efficiency in automated scoring (AS) of reading assessments, which typically involve lengthy passages and complex scoring guides. We propose optimizing input size using question-specific summaries and simplified scoring guides. Findings indicate that input optimization via compression is achievable while maintaining AS performance.

1 Introduction

Automated scoring (AS) has a rich history in educational measurement (Lottridge et al., 2023), dating back to the 1960s when the primary focus was on scoring multiple-choice responses or implementing machine-supported scoring based on pattern matching or manual feature selection. The rapid advances in natural language processing (NLP), machine learning, and computational power have led to significant developments in large language models (LLMs). Integrating LLMs, such as OpenAI's GPT models or META's Llama, into AS expands the applicability and scalability of AS in educational assessment.

However, applying LLMs to the AS of reading assessments presents unique challenges in processing long inputs, including extended reading passages and complex scoring guides (SGs). Given that the cost of using LLMs through APIs depends on the number of input, cached, and output tokens (OpenAI, 2025), extensively long prompts can lead to inflated costs for each API call. Moreover, previous study indicated that long prompts can cause a "lost in the middle" effect, where LLMs struggle to appropriately use the most relevant context embedded within the extensive input (Liu et al., 2023). This limitation persists, particularly for smaller models operated locally.

To address the challenge of processing long inputs, we propose input optimization to improve the scalability and efficiency of AS in international large-scale assessments (ILSAs).

2 Background

Very long inputs can slow LLMs' inference processes and increase energy use due to the increased the number of tokens that need to be processed. Prior research showed that LLMs do not robustly utilize information in long input contexts and may ignore parts of the given context, generating incorrect outputs (Liu et al., 2023). Crucially, extended input lengths lead to a linear increase in both computational costs and energy demands (Poddar et al., 2025).

Text compression shrinks textual data while preserving crucial information, improving storage and computational efficiency, and enhancing the performance of LLMs (Rahman et al., 2024; Wang et al., 2024). Compression can be achieved through either soft or hard prompts. Soft prompts are continuous vectors, enabling LLMs to address long and complex input by distilling critical information into a smaller number of special tokens (Li et al., 2024; Wang et al., 2024). Yet, soft prompts are less interpretable by humans and are often highly customized to specific tasks. Their reusability or transferability across different tasks can be constrained (Su et al., 2022).

In contrast, hard prompts comprise discrete words and tokens, making them easily understandable by humans. This readability and transparency allow humans to review, debug, and modify prompts by facilitating effective human-machine interaction (Chang et al., 2024; Wen et al., 2023). Hard prompts can be especially powerful when prompts need human interpretation or are integrated into a text-based interface (Wen et al., 2023; Jiang et al., 2023). Zhang et al. (2024) found that hard prompts yield superior performance for

Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con) – Volume 1: Full Papers, pages 1–8

summarization compared to soft prompts in human evaluations.

Despite the demonstrated usefulness of text compression techniques, they have not been widely integrated into AS for reading assessments in ILSAs, such as the Progress in International Reading Literacy Study (PIRLS). Optimizing long input through compression in reading assessments can contribute to improving AS scalability and cost- and computational efficiency in ILSAs. This paper examines how advances in hard prompt-based input optimization can be integrated into AS in PIRLS, which involves a substantial volume of multilingual responses.

3 Method

3.1 Dataset

The PIRLS, administered every five years since 2001, assesses the reading comprehension skills of fourth-grade students across 50-60 countries worldwide. In PIRLS 2021, approximately 50% of countries (27 countries) used computer-based assessments. The assessment framework categorizes reading comprehension into four cognitive processes: focus on and retrieve; straightforward inferences; interpret and integrate; and evaluate and critique (Mullis & Martin, 2019). For this study, we selected five one-point constructed response (CR) items from the PRILS 2021 digital assessment (digital PIRLS). The selected items represent three cognitive processes: one from focus on and retrieve, two from straightforward inferences, and two from interpret and integrate.

These items are "trend" items, kept secure for their reuse in future assessment cycles (Fishbein et al., 2024). We provide general descriptions of these items (Table 1) as this research is part of the preparatory work for AS in PIRLS 2026, where these items will be used. We selected four reading passages with varying difficulty levels: easy (passages B and D), medium (passage A), and difficult (passage C).

Item	Passage	Process	*n*
1	A	Focus on and retrieve	2687
2	B	Straightforward inferences	2951
3	A	Straightforward inferences	2643
4	C	Interpret and integrate	2589
5	D	Interpret and integrate	2452

Table 1: PIRLS trend items used in the study

The dataset included multilingual responses from the 27 participating countries in digital PIRLS 2021, covering 29 languages. While approximately 50% of participating countries used computer-based assessments in PIRLS 2021, the data still contained on average, 2,664 multilingual responses per item (see Appendix A). We used a randomly selected 20% subset for each country given the scope, computational and budgetary limitations.

3.2 PIRLS Scoring Template

We proposed a generalized PIRLS scoring template for AS (see Appendix B), comprising four key elements: (1) instruction, (2) reading passage, (3) question, and (4) SG, as detailed in Table 2. We used GPT-4.1 (i.e., gpt-4.1-2025-04-14) for our AS implementation, applying parallel processing for efficiency. This template used zero-shot chain-of-thought (CoT), a technique that enhances LLM performance through step-by-step reasoning without requiring specific examples (Kojima et al., 2022; Yuan et al., 2024). Zero-shot COT offers the advantage of easy generalization to other items due to its independence from specific examples.

Component	Content
Instruction	Comprehensive guidance on AS
Reading passage	A written text serving as the stimulus
Question	A question consisting of one or two sentences
Scoring guide (SG)	Rubric for scoring an item, including descriptions and examples

Table 2: PIRLS scoring template components

Instruction: The instruction component offers comprehensive guidance on translating student responses, applying the SG, validating scores, and constructing output.

Reading Passage: The second component, a reading passage, could be presented as either the original passage or a question-specific summary. Original passages provide the complete

information as presented to students, whereas summaries include question-relevant details while preserving overall context.

Question: The third component, the item's question, was directly input into the scoring template.

Scoring Guide (SG): The SG could be either the original SG or a simplified version. Simplified SGs were designed to mitigate challenges arising from ambiguous structure or meaning in the original SGs, which may lead to less accurate output from LLMs. Prior studies (Keluskar & Bhattacharjee, 2024; Kamath et al., 2024) indicate that rephrasing or clarifying sentences in prompts can significantly improve LLM output quality.

3.3 Input Optimization

Question-specific Summary: The passage summarization prompt shown in Figure 1 was used to generate question-specific summaries that retain all essential information needed to answer the question while maintaining the overall flow. Query-based text summarization aids users in accessing specific information within lengthy texts, enabling LLMs to provide efficient access to relevant content (Yu & Han, 2022; Zhang et al., 2025). This zero-shot CoT prompt can be applied across various items, requiring only the [[question]] input to be modified.

Summarize the passage for a fourth-grade student, including the overall flow and all necessary information to correctly answer the question: [[question]]

1. Read the Passage: Carefully read the passage to understand the main events and details.
2. Summarize: Create a summary that includes the overall flow, and the necessary information related to the question.
3. Final Output
   - The output should be a coherent paragraph summarizing the passage.
   - Avoid new section headings.

Figure 1: Passage summarization prompt

Simplified SG: Original SGs for one-point items in PIRLS 2021 consist of two parts: a description with examples of acceptable responses, and a description with examples of unacceptable responses. For the simplified SG, we utilized GPT-4.1 to improve the readability of acceptable response descriptions from the original SGs. This involved rephrasing or reconstructing sentences and removing examples, guided by the SG modification prompt (Figure 2). For unacceptable response descriptions, we adopted a standard description: "Assign this score if the response does not explicitly include the key content described in the [Score: 1] criteria." Replacing the original item-specific descriptions.

Additionally, we incorporated notes reflecting the general guidelines of the PIRLS Scoring Guides: "(1) Minor irrelevant details are permissible only if the response explicitly includes the key content required for [Score: 1] and the details do not contradict the [Score: 1] criteria. (2) Character names may vary depending on the language used; such variations should not affect scoring."

Improve the language in the current scoring guide.

Steps
1. Review the Scoring Guide: Carefully read the existing scoring guide to grasp its content and scoring criteria.
2. Refine Language: Enhance the language for clarity while keeping the intended meaning of the original scoring guide.
3. Final Output: Produce the final output in plain text.

Output Format
- Use bullet points if they improve readability.
- Maintain the given structure: "**[Score: X]**: Assign this score if …"
- Avoid new section headings or providing examples.

Figure 2: SG modification prompt

3.4 AS with PIRLS Scoring Template

We ran two separate AS models using the PIRLS scoring template: a baseline model and an optimized AS model with compression (Opt-AS). The baseline model used the original reading passages and SGs, while the Opt-AS model integrated question-specific summaries and simplified SGs. For each item, a single summary and simplified SG were created and consistently applied to all responses. Following the Opt-AS, custom Python scripts were utilized to

automatically identify and correct mis-formatted outputs to ensure a consistent format.

3.5 Evaluation Metrics

We evaluated AS performance using four metrics: compression ratio, exact agreement (EA), and Cohen's Kappa (κ).

Compression ratio quantifies the efficiency of our input optimization by comparing the token count of optimized inputs to that of original inputs. We specifically focused on the token reduction in reading passages and SGs, where lower values indicate higher compression. For SGs, notes reflecting the general guidelines of the PIRLS Scoring Guides were excluded from the compression ratio calculation.

$$R = \frac{Token\ count\ of\ optimized\ input}{Token\ count\ of\ original\ input} \quad (1)$$

EA, a commonly used metric in AS, is calculated as the percentage of exact matches between human and machine scores.

Cohen's Kappa (Cohen, 1960) measures inter-rater reliability by considering chance agreement, and is calculated as follows:

$$\kappa = \frac{p_o - p_e}{1 - p_e} \quad (2)$$

where p_o is the observed agreement among raters, and p_e denotes the expected probability of chance agreement. The Kappa ranges from 0 (agreement due to chance) to 1 (perfect agreement).

We computed processing time and estimated costs for Opt-AS using Python scripts. Cost estimates were based on the number of input and output tokens, following the GPT-4.1 API pricing (OpenAI, n.d.): \$2.00 per million input tokens and \$0.80 per million output tokens. One million tokens are approximately equivalent to 750,000 words.

4 Results

Compression Ratio: Tables 3 and 4 present token counts and compression ratios. On average, passages were compressed to 20.22% of the original length, while SGs were reduced to 46.47% of their original size.

	Baseline		Opt-AS	
Item	Passage	SG	Passage	SG
1	724	112	168	67
2	581	119	117	93
3	724	152	155	65
4	1045	163	168	79
5	640	261	143	71
Avg.	743	161	150	75

Table 3: Token count for passage and SG

Item	Passage	SG
1	23.20%	59.82%
2	20.14%	78.15%
3	21.41%	42.76%
4	16.08%	48.47%
5	22.34%	27.20%
Avg.	20.22%	46.47%

Table 4: Compression ratio

EA & Kappa: Our Opt-AS model demonstrated comparable performance to the baseline model, achieving an average EA of 95.16% and kappa of 0.8852. Notably, for Item 1, the Opt-AS model yielded a lower kappa of 0.8482 compared to the baseline (0.9308). This discrepancy can be attributed to Item 1 being a very easy item, resulting in highly imbalanced data where 91.9% of responses received a human score of 1. Despite this, Opt-AS maintained strong precision and recall values of 98.55% and 98.34%, respectively (see confusion matrices in Appendix C).

Item	Baseline		Opt-AS	
	EA	κ	EA	κ
1	98.78%	0.9308	97.18%	0.8482
2	96.13%	0.9203	96.35%	0.9231
3	94.35%	0.8609	94.47%	0.8750
4	93.64%	0.8706	93.48%	0.8768
5	93.27%	0.8570	93.50%	0.8511
Avg.	95.16%	0.8852	94.94%	0.8723

Table 5: EA & Kappa

Processing Time & Cost: The average processing time and cost per item using Opt-AS were approximately 6 minutes and \$3.09, respectively (see Table 6). In contrast to the extensive resources required for human rater training and scoring (Ward & Bennett, 2012), this

reflects a highly efficient use of time and cost. Moreover, our Opt-AS reduced costs by nearly 50% relative to the baseline model, which incurred approximately $6 per item and required around 7 minutes of processing time.

Item	Processing Time	Cost ($)
1	00:06:05	3.170
2	00:07:17	3.390
3	00:05:59	2.755
4	00:06:32	3.210
5	00:06:19	2.907
Avg.	00:06:26	3.087

Table 6: Processing time & cost

5 Discussion

Our findings indicate that input optimization significantly reduces the complexity of AS in reading assessments. Aligned with prior research (Jiang et al., 2023; Xu & Lapata, 2022), Opt-AS leverages compression techniques to optimize input size, substantially shortening text length while preserving critical information. This optimization effectively lowers computational costs without compromising AS performance, even on low-resource languages such as Arabic, Croatian, and Maltese. Given the considerable cost and time involved in scoring over 12,000 multilingual written responses per CR item in PIRLS, and the shift to fully digital assessment for all participating countries in PIRLS 2026 (von Davier & Kennedy, 2024), Opt-AS offers a cost-effective, energy-efficient, and scalable scoring solution in a computer-based assessment context.

Despite these promising results, this study has limitations. First, due to its exploratory nature, the analysis was conducted on a randomly selected 20% sample. While this sample was representative, future research should assess the generalizability of our approach using the full PIRLS dataset across a broader range of CR items. Next, further investigation into AS consistency is necessary. Although GPT-4.1's temperature was set to 0 to minimize variability, validating the consistency of both AS and human scoring remains important. One potential method is to use sentence embedding techniques to cluster semantically similar responses, allowing for a systematic evaluation of scoring consistency across both scoring methods.

6 Conclusion

This study provides compelling evidence for the effectiveness of input optimization for AS in multilingual reading assessments. Our Opt-AS approach maintained robust performance within the PIRLS framework, concurrently saving time, cost, and computational burden. The streamlined AS enhances operational efficiency and scalability across a multitude of assessment items and countries. Ultimately, well-implemented AS systems promise to deliver timely, accurate, and reliable reporting to participating countries, supporting more informed educational policy decisions.

References

Chang, K., Xu, S., Wang, C., Luo, Y., Liu, X., Xiao, T., & Zhu, J. (2024). Efficient prompting methods for large language models: A survey. arXiv preprint arXiv:2404.01077.

Cohen, J. (1960). A coefficient of agreement for nominal scales. Educational and psychological measurement, 20(1), 37-46.

Fishbein, B., Yin, L., & Foy, P. (2024). PIRLS 2021 User Guide for the International Database (2nd ed.). Boston College, TIMSS & PIRLS International Study Center. https://pirls2021.org/data

Jiang, H., Wu, Q., Lin, C. Y., Yang, Y., & Qiu, L. (2023). Llmlingua: Compressing prompts for accelerated inference of large language models. arXiv preprint arXiv:2310.05736.

Kamath, G., Schuster, S., Vajjala, S., & Reddy, S. (2024). Scope ambiguities in large language models. Transactions of the Association for Computational Linguistics, 12, 738-754.

Keluskar, A., Bhattacharjee, A., & Liu, H. (2024, December). Do LLMs Understand Ambiguity in Text? A Case Study in Open-world Question Answering. In 2024 IEEE International Conference on Big Data (BigData) (pp. 7485-7490). IEEE.

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2022). Large language models are zero-shot reasoners. Advances in neural information processing systems, 35, 22199-22213.

Li, Z., Liu, Y., Su, Y., & Collier, N. (2024). Prompt compression for large language models: A survey. arXiv preprint arXiv:2410.12388.

Liu, N. F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., & Liang, P. (2023). Lost in the middle: How language models use long contexts. arXiv preprint arXiv:2307.03172.

Lottridge, S., Ormerod, C., & Jafari, A. (2023). Psychometric considerations when using deep learning for automated scoring. Advancing natural language processing in educational assessment, 15.

Mullis, I. V. S., & Martin, M. O. (Eds.). (2019). PIRLS 2021 Assessment Frameworks. Retrieved from Boston College, TIMSS & PIRLS International Study Center website: https://timssandpirls.bc.edu/pirls2021/frameworks/

OpenAI. (n.d.). API pricing. Retrieved May 16, 2025 from https://openai.com/api/pricing/

Poddar, S., Koley, P., Misra, J., Podder, S., Ganguly, N., & Ghosh, S. (2025). Towards Sustainable NLP: Insights from Benchmarking Inference Energy in Large Language Models. arXiv preprint arXiv:2502.05610.

Rahman, C. M., Sobhani, M. E., Rodela, A. T., & Shatabda, S. (2024, September). An Enhanced Text Compression Approach Using Transformer-based Language Models. In 2024 IEEE Region 10 Symposium (TENSYMP) (pp. 1-6). IEEE.

Su, Y., Wang, X., Qin, Y., Chan, C. M., Lin, Y., Wang, H., ... & Zhou, J. (2021). On transferability of prompt tuning for natural language processing. arXiv preprint arXiv:2111.06719.

von Davier, M. & Kennedy, A., Editors (2024). PIRLS 2026 Assessment Frameworks. Boston College, TIMSS & PIRLS International Study Center https://doi.org/10.6017/lse.tpisc.tr2103.kb4199

Wang, C., Yang, Y., Li, R., Sun, D., Cai, R., Zhang, Y., & Fu, C. (2024, May). Adapting llms for efficient context processing through soft prompt compression. In Proceedings of the International Conference on Modeling, Natural Language Processing and Machine Learning (pp. 91-97).

Ward, W. C., & Bennett, R. E. (2012). Construction versus choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment. Routledge.

Wen, Y., Jain, N., Kirchenbauer, J., Goldblum, M., Geiping, J., & Goldstein, T. (2023). Hard prompts made easy: Gradient-based discrete optimization for prompt tuning and discovery. Advances in Neural Information Processing Systems, 36, 51008-51025.

Xu, Y., & Lapata, M. (2022). Document summarization with latent queries. Transactions of the Association for Computational Linguistics, 10, 623-638.

Yuan, X., Shen, C., Yan, S., Zhang, X., Xie, L., Wang, W., ... & Ye, J. (2024). Instance-adaptive zero-shot chain-of-thought prompting. arXiv preprint arXiv:2409.20441.

Zhang, W., Huang, J. H., Vakulenko, S., Xu, Y., Rajapakse, T., & Kanoulas, E. (2025). Beyond relevant documents: A knowledge-intensive approach for query-focused summarization using large language models. In International Conference on Pattern Recognition (pp. 89-104). Springer, Cham.

Zhang, Y., Liu, Y., Yang, Z., Fang, Y., Chen, Y., Radev, D., ... & Zhang, R. (2023). Macsum: Controllable summarization with mixed attributes. Transactions of the Association for Computational Linguistics, 11, 787-803.

A Appendices

A. Sample Size by Country

Country	Item 1	Item 2	Item 3	Item 4	Item 5
A	410	524	406	342	449
B	76	82	77	73	68
C	226	252	219	230	212
D	111	119	107	104	100
E	69	70	67	n/a	56
F	72	74	69	64	58
G	126	138	121	127	120
H	102	112	100	142	82
I	60	58	61	60	47
J	80	89	79	79	75
K	85	90	84	69	77
L	107	118	107	99	100
M	67	67	67	63	52
N	46	46	45	43	45
O	80	88	79	79	72
P	83	90	82	76	77
Q	93	92	90	90	79
R	78	87	79	76	70
S	86	91	86	119	77
T	77	82	77	109	70
U	75	80	74	39	64
V	100	104	99	137	80
W	70	73	69	63	57
X	75	79	73	76	63
Y	76	81	76	76	66
Z	82	89	79	79	76
AA	75	76	71	75	60
Total	2687	2951	2643	2589	2452

Table. Sample size by country

B. PIRLS Scoring Template

Evaluate multilingual responses from an international reading assessment for fourth-grade students.

Steps

1. Translation: Translate the student's response into English.
2. Scoring: Score the response according to the given scoring guide.
3. Validation: Determine if the translation could be "hallucinated" where the text appears linguistically correct but fails to capture the intended meaning.
    - If the translation is inaccurate, re-translate and re-score the response.
    - If the original text is untranslatable and nonsensical, keep the original text and assign a score of 0.
4. Output Construction: Compile the result into a JSON object, with either the translated text or the original text (if untranslatable) and the assigned score.

Output Format

The output should be formatted in JSON as follows: {"[English translation or original text]": "Score: [score]"}

Passage: **[[Original reading passage or question-specific summary]]**

Question: **[[Item's question]]**

Scoring Guide:

Evaluate responses based on the following criteria.

- [Score: 1]: Assign this score if **[[description]]**
- [Score: 0]: Assign this score if the response does not explicitly include the key content described in the [Score: 1] criteria.

Notes

- Minor irrelevant details are permissible only if the response explicitly includes the key content required for [Score: 1] and the details do not contradict the [Score: 1] criteria.
- Character names may vary depending on the language used; such variations should not affect scoring.

C. Confusion Matrices from Optimized AS

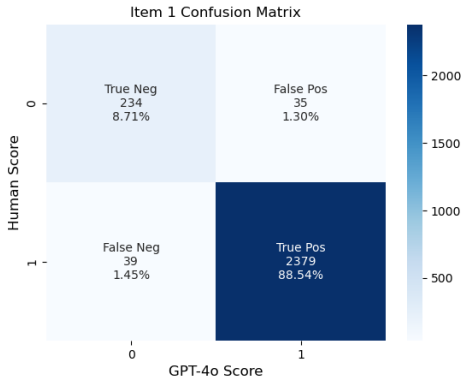

Figure 1. Item 1 confusion matrix

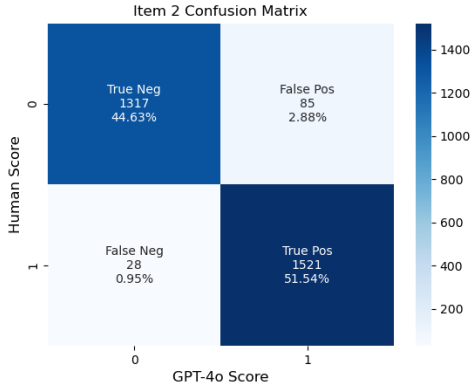

Figure 2. Item 2 confusion matrix

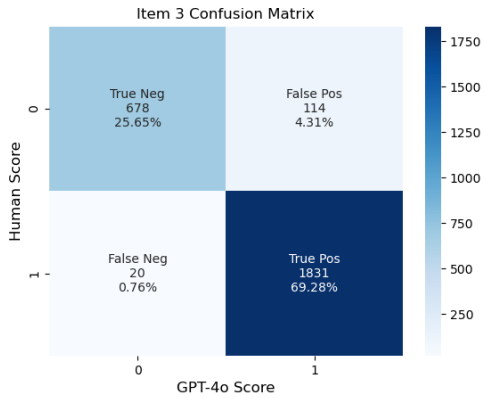

Figure 3. Item 3 confusion matrix

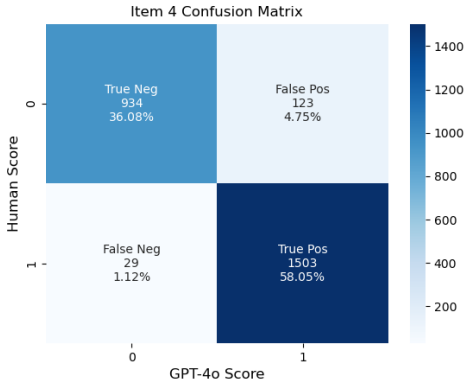

Figure 4. Item 4 confusion matrix

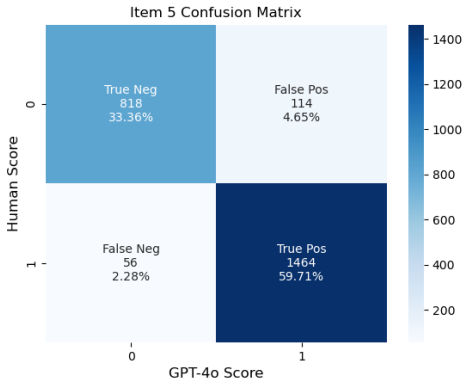

Figure 5. Item 5 confusion matrix