

Comprehensive Pipeline for Multilingual GenAI Scoring

Ji Yoon Jung Ummugul Bezirhan Matthias von Davier

TIMSS & PIRLS International Study Center at Boston College

{jungjg, bezirhan, vondavim}@bc.edu

Abstract

This study presents a comprehensive operational pipeline for multilingual GenAI scoring that integrates secure data preparation, prompt generation, automated scoring, post-processing, and human-in-the-loop review. Results demonstrate the pipeline's strong adaptability across item formats, assessment languages, and assessment cycles, offering a viable solution to the persistent challenges of multilingual scoring.

1 Introduction

Automated scoring (AS) is experiencing transformative possibilities with the rapid advances in generative artificial intelligence (GenAI), particularly large language models (LLMs) such as GPT and Gemini. These innovations significantly enhance cost-efficiency and scalability, offering substantially lower scoring costs per item compared to human scoring. Recent research (Frohn et al., 2025; Jung et al., 2025c) has further supported this shift, demonstrating that GenAI scoring achieves performance comparable to human scoring while remaining highly cost-effective. These advantages are particularly valuable for international large-scale assessments (ILSAs), which typically demand massive volume scoring across more than 60 languages and often at least twice as many language versions (variants of the same language, with differences that are country- or region-dependent).

Despite these technological gains, a comprehensive operational workflow for multilingual GenAI scoring remains to be established. An end-to-end pipeline is essential to ensure security, reproducibility, and scalability alongside reliability, validity, and comparability across item formats, languages spoken in countries, and assessment cycles. Previous research (Latif &

Zhai, 2024; Lee et al., 2024) has focused on specific facets, such as prompt engineering, hyperparameter tuning, or performance comparability, addressing individual elements rather than the complete pipeline required for large-scale implementations. While some studies (Shin et al., 2024; Okubo et al., 2023; Zehner et al., 2025) have proposed operational frameworks, they remain tied to supervised learning, necessitating extensive feature engineering and high-quality training datasets. Therefore, there remains a critical need for a comprehensive framework designed for the unique flexibility and requirements of multilingual GenAI scoring.

To address this gap, this paper introduces Compass (COMprehensive Pipeline for AI Scoring System). Compass, to our knowledge, the first operational multilingual GenAI scoring framework that unifies PII protection, automated prompt generation, hallucination mitigation, consistency validation, and expert review into a single workflow. Using data collected with constructed-response (CR) items from the Progress in International Reading Literacy Study (PIRLS), this study evaluates Compass' scoring performance across diverse items and countries.

2 Related Work

AS systems have evolved from traditional feature-based architectures, such as ETS's e-rater and Pearson's Intelligent Essay Assessor, to modern supervised learning engines. While legacy engines rely on rigorous practices that involve explicit feature extraction and monitoring (McCaffrey et al., 2022), contemporary models like Duolingo's scoring engine leverage transformer-based architectures and deep learning to achieve greater alignment with human scoring (Zehner et al., 2025; McCaffrey et al., 2022; Naismith et al., 2025). However, established processes for supervised

models are not directly transferable to GenAI scoring. The shift toward GenAI, which features prompt-based inference, bypasses the explicit feature extraction used in legacy systems, requiring a fundamental reimagining of operational frameworks.

Integrating GenAI into operational scoring systems requires addressing unique practical challenges. First, data privacy poses a primary concern, as the use of GenAI typically requires transmitting student responses to external APIs. This workflow introduces potential exposure risks for personally identifiable information (PII) such as names, or email addresses (Kaddour et al., 2023) that students may have entered instead of, or in addition to, a response to the question at hand.

Beyond privacy, GenAI outputs are highly sensitive to prompt design; both the structure and the content of instructions substantially influence model performance (Chen et al., 2025; Ekin, 2023). In the AS context, Mansour et al. (2024) demonstrated that model specific prompting is vital for optimizing scoring accuracy. Moreover, prompt design is central for building validity evidence. Prompts for AS need to provide analogous information to what human scorers receive as instruction. That typically means the prompts need to be constructed to include all relevant information about the item and how to score responses, including information from scoring guides. The materials describing how to score responses provided to human raters need to be incorporated into the prompt, as they operationalize how scores assigned to responses should align with the construct to be measured and define the criteria for response evaluation. If a prompt poorly specifies this essential information, the model may apply criteria that diverge from the content experts' intent, thereby undermining validity evidence.

Furthermore, the tendency of GenAI to produce confident yet nonfactual outputs, or hallucinations, underscores the need for rigorous external validation (Gunjal et al., 2024). Ultimately, as Thomas et al. (2026) emphasize, the responsible integration of GenAI requires a robust, transparent framework to ensure reliable performance.

3 Method

3.1 Dataset

PIRLS has assessed reading comprehension of fourth-grade students globally every five years since 2001 using passage-based testlets that ask students to read a story and answer questions about it. In the present study, we utilized six one-point CR items drawn from two cycles: digital PIRLS 2021, adopted by approximately half of participating countries (27 countries, 29 languages), and the PIRLS 2026 field test (56 countries, 40 languages). Item details are provided in Table 1. The items were taken from the item sets associated with six different reading passages.

Item	Cycle	Process	<i>n</i>
1	2021	Focus on & retrieve	14,087
2	2021	Straightforward inferences	13,138
3	2021	Straightforward inferences	14,718
4	2026	Focus on & retrieve	14,622
5	2026	Straightforward inferences	13,335
6	2026	Straightforward inferences	14,104

Table 1: PIRLS items used in this study.

3.2 Compass Pipeline

Figure 1 illustrates the overall processing pipeline for Compass. On the left side, the inputs (i.e., raw responses, reading passage, and scoring guide) are shown in blue text boxes. Moving to the right, the figure depicts data preparation and automated prompt generation; automated scoring; automated post-processing, and human expert review, with the final score as the output on the right.

PII Masking: The data preparation process begins with multilingual PII masking to ensure data privacy, which leverages Microsoft's Presidio for structured PII and the XLM-RoBERTa model (Liu et al., 2019) for identifying multilingual proper nouns. Legitimate proper nouns appearing in the reading passages (e.g., character or location names) were not treated as PII, as students needed to use these terms in their responses. Specifically, our system categorizes languages into Arabic, Cyrillic (e.g., Russian), Latin (e.g., English, French), Chinese, and Korean groups. For each category, we

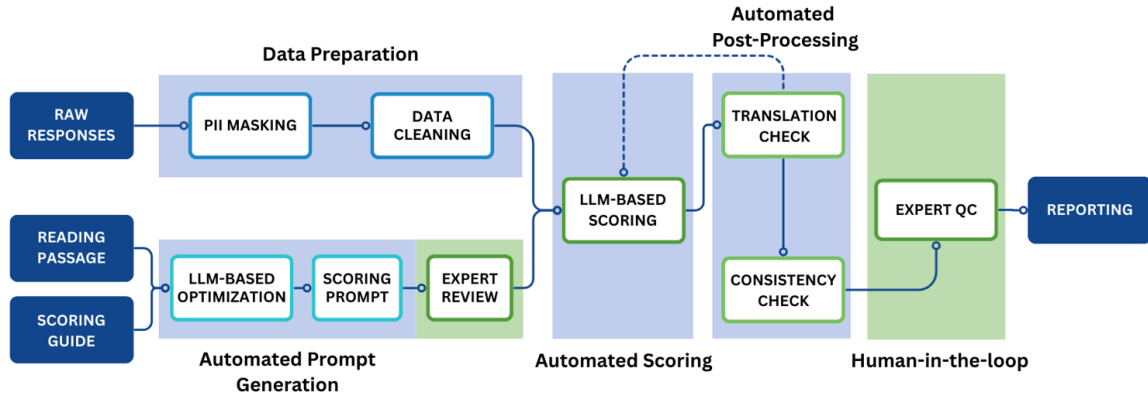


Figure 1: Comprehensive scoring pipeline.

compiled the 20 most frequent personal names and generated allowlists containing their orthographic and phonetic variants via Levenshtein distance (Levenshtein, 1966) and Soundex (Odell & Russell, 1918). These allowlisted variants were bypassed during final PII detection to prevent valid, passage-specific names from being incorrectly flagged, even if misspelled. Once sensitive content was masked, the data was cleaned by standardizing whitespace and line breaks and removing artifacts such as special characters and excessive repetitions.

Scoring Prompt Generation: Scoring prompts were automatically generated using a standard template built upon a chain-of-thought (CoT) technique (Jung et al., 2025a; Kojima et al., 2022). This template, shown in Appendix A, comprised four components: task instructions, reading passage, question, and scoring guides. This core instruction was to score multilingual responses while also providing an English translation as a means of quality control for non-English responses. This translation serves not as a prerequisite for scoring, but rather as a CoT step designed to facilitate the model’s reasoning process. Crucially, to improve GenAI efficiency, the template utilized question-specific passage summaries and simplified scoring guides, rather than the original, lengthy versions (Jung et al., 2025a). These design choices reduced prompt length while preserving construct-relevant information necessary for valid scoring. Scoring prompts were validated and finalized by content experts.

AI Scoring: Scoring utilized the GPT4 API, processing PII-masked responses along with their associated scoring prompts. Each response was tagged with its source language to achieve accurate

scoring and translation. Importantly, this tagging mechanism prevents cross-lingual inputs from contaminating the multilingual scoring process. Low-effort responses (e.g., comprising only a single punctuation mark, letter, or number) were preemptively scored as zero, reducing API calls. The API temperature was set to 0 to ensure highly consistent and deterministic scoring. After scoring, the machine-generated English translations were stored, along with API costs and processing time for tracking.

Automated Post-Processing: To ensure translation quality, we performed hallucination detection to capture multiple patterns, including translation truncation, severe semantic divergence, and more subtle cases involving abnormal lengths combined with semantic drift. Initial detection relied on two metrics: the length ratio (translation-to-original text) and the cosine similarity between the original and translated text, derived from LaBSE multilingual embeddings. A combined metric incorporating both was further used to identify mixed-abnormality cases where length and semantic draft issues co-occurred. Thresholds for all metrics were determined empirically through iterative inspection of translation outputs across languages, balancing sensitivity to hallucinations against the risk of over-flagging valid paraphrase translations.

Next, the original texts of the flagged translations were cross-referenced with those in the non-hallucinated dataset. Using fuzzy matching and edit distance, we identified identical or nearly identical original texts to find the closest match for each flagged entry. If non-hallucinated responses using fuzzy matching (ratio > 0.85) or edit distance (< 3); if a match was found, the flagged translation was replaced with the corresponding verified

alternative. Any responses remaining flagged after this substitution stage underwent API re-scoring. If a translation remained flagged after this API re-scoring, the response was excluded from the final AS and flagged for expert review.

The consistency check served to cluster lexically similar responses and standardized their assigned scores. As this process operated on the translated text, the preceding hallucination check was essential to ensure the translation quality. The DBSCAN clustering algorithm (Ester et al., 1996) was applied using edit distance as the metric, grouping responses with an edit distance of 1 and requiring a minimum of 3 neighbors per cluster. Within each resulting cluster, if a single score accounted for more than 80% of the responses, that majority score was applied to all members of that cluster to enhance score consistency.

Results were saved at two levels: response-level data, containing original/translated text and human/machine scores; and aggregate-level, summarizing metrics such as exact agreement (EA) and Cohen’s kappa (κ). These metrics were aggregated by country, language, and language within country.

Expert Quality Control: The AS development team and content experts performed targeted quality control. This involved identifying and reviewing systematic inconsistencies or ambiguous

Item	$EA_{initial}$	EA_{final}	$\kappa_{initial}$	κ_{final}
1	95.46	95.63	0.91	0.91
2	93.96	94.00	0.78	0.78
3	95.87	95.92	0.88	0.88
4	98.04	98.06	0.90	0.90
5	95.00	95.06	0.90	0.90
6	96.21	96.24	0.91	0.91
Avg.	95.76	95.82	0.88	0.88

Table 2: Scoring performance for PIRLS items.

responses. Also, country- and language-specific anomalies were examined to diagnose and address potential underlying scoring or administrative issues.

4 Results

4.1 Performance Evaluation

Compass achieved exceptionally high agreement with human scoring across all items ($EA_{final} = 95.82$, $\kappa_{final} = 0.881$). Post-processing algorithms provided an additional layer of quality assurance, reducing translation errors and improving scoring

consistency, which yielded small but consistent gains in agreement with human scoring across all items (see Table 2).

Figure 2 illustrates the cross-country robustness

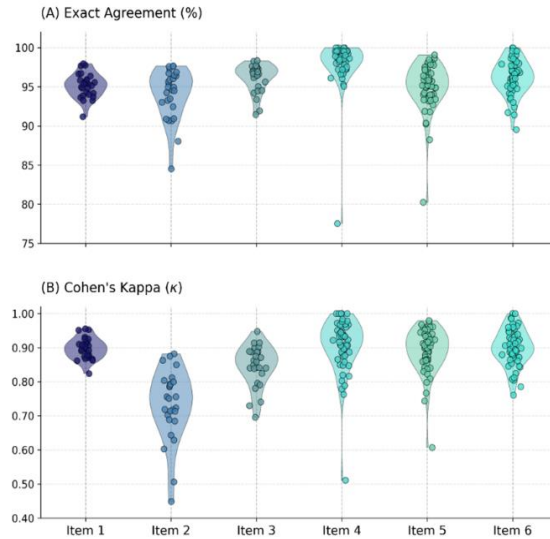


Figure 2: Compass scoring performance.

of Compass, demonstrating consistently high EA_{final} and κ_{final} across all items and countries. For most countries, EA_{final} consistently exceeded 90%, and κ_{final} generally remained above 0.70. Yet, one country showed outlying performance on Item 4, and another on Item 5. Content expert reviews of these cases are detailed in the Discussion.

4.2 Efficacy of Automated Post-Processing

The translation check proved effective in reducing hallucinations. While the initial flagging identified an average of 104 hallucinated responses (0.74%), the subsequent matching algorithm and GPT-based rescoring reduced this count to 27 cases, representing only 0.19% of the total sample. Most of the remaining hallucinated responses originated from low-resource languages, where GenAI have limited language-specific training data. English responses containing severe orthographic issues, such as extensive misspellings, grammatical errors, or missing spacing between words, also contributed to the residual hallucinations.

The consistency check also enhanced scoring reliability through cluster-based validation. The algorithm created an average of 787 clusters per item, and in 97.63% of these clusters, the majority score was uniformly assigned to all members. On average, the scores of 15 responses per item were

updated to align with the majority scores. While minimal in scale, these highly targeted corrections contributed to the observed improvements in overall scoring performance ($\Delta EA = +0.06$, $\Delta \kappa = 0.001$). Moreover, these clusters provided diagnostic insights, enabling content experts to efficiently identify challenging response patterns instead of reviewing individual responses.

5 Discussion and Future Work

We introduce Compass, a holistic pipeline for GenAI-based AS. While prior work largely relied on supervised learning or focused on individual applications of GenAI, Compass offers a transparent, operational workflow for multilingual scoring. By integrating secure data preparation, automated prompt generation, scoring, post-processing, and expert review, the pipeline addresses practical challenges associated with large-scale implementation while maintaining scoring quality. Crucially, Compass integrates automated postprocessing checks for hallucination and score consistency, followed by content expert review. This structured approach makes economical and reliable multilingual scoring attainable, addressing long-standing barriers in ILSAs (Hao et al., 2024).

Although the post-processing procedures yielded only modest improvements in overall agreement metrics, they served an important quality assurance function. In particular, the hallucination detection workflow reduced potentially problematic translations from 0.74% to 0.19% of responses, while the consistency check identified and corrected a small number of scoring discrepancies among highly similar responses. These findings suggest that post-processing contributes not only to scoring accuracy but also to the transparency and trustworthiness of GenAI-based scoring systems.

Despite strong performance, instances of lower agreement occurred in specific contexts, such as Item 4 in Country 1 ($EA_{final} = 77.84\%$) and Item 5 in Country 2 ($EA_{final} = 80.25\%$). Expert review revealed that this variability stems largely from highly nuanced responses posing inherent scoring challenges. For instance, human scoring in Country 1 showed general inconsistency, whereas human scoring in Country 2 showed greater leniency toward specific borderline responses than in other countries. Lower agreement in these cases can be attributed to the inherent complexity of items that

require fine-grained distinctions and to divergent interpretations of responses that fall on the scoring boundary.

This study has two primary limitations. First, although validated on the PIRLS dataset, the pipeline's generalizability to domains such as mathematics or science requires further confirmation. Second, as the consistency check relies on dense lexical clustering, unique responses fall outside this highly targeted validation scope. Future work can explore semantic similarity-based grouping to extend these checks. For example, we plan to integrate the Linguistic-integrated Reliability Audit (LiRA) framework (Jung et al., 2025b) to automatically generate response-level audit scores, enhancing internal reliability of AS beyond the current lexical approach.

References

- Chen, B., Zhang, Z., Langrené, N., & Zhu, S. (2025). Unleashing the potential of prompt engineering for large language models. *Patterns*, 6(6).
- Ekin, S. (2026). Prompt engineering for ChatGPT: a quick guide to techniques, tips, and best practices. *Journal of Engineering Technology*, 43(1).
- Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996, August). A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd* (Vol. 96, No. 34, pp. 226-231).
- Frohn, S., Burleigh, T., & Chen, J. (2025, July). Automated Scoring of Short Answer Questions with Large Language Models: Impacts of Model, Item, and Rubric Design. In *International Conference on Artificial Intelligence in Education* (pp. 44-51). Cham: Springer Nature Switzerland.
- Gunjal, A., Yin, J., & Bas, E. (2024, March). Detecting and preventing hallucinations in large vision language models. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 38, No. 16, pp. 18135-18143).
- Hao, J., von Davier, A. A., Yaneva, V., Lottridge, S., von Davier, M., & Harris, D. J. (2024). Transforming assessment: The impacts and implications of large language models and generative AI. *Educational Measurement: Issues and Practice*, 43(2), 16-29.
- Jung, J. Y., Bezirhan, U., & von Davier, M. (2025a). Input optimization for automated scoring in reading assessment. In *Proceedings of the artificial intelligence in measurement and education conference (AIME-con): Full papers* (pp. 1-8).
- Jung, J. Y., Bezirhan, U., & von Davier, M. (2025b). Reconceptualizing Scoring Reliability Through

- Linguistic Similarity. *Educational and Psychological Measurement*, 00131644251397428.
- Jung, J. Y., Tyack, L., & von Davier, M. (2025c). Towards the implementation of automated scoring in international large-scale assessments: Scalability and quality control. *Computers and Education: Artificial Intelligence*, 8, 100375.
- Kaddour, J., Harris, J., Mozes, M., Bradley, H., Raileanu, R., & McHardy, R. (2023). Challenges and applications of large language models. *arXiv preprint arXiv:2307.10169*.
- 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.
- Latif, E., & Zhai, X. (2024). Fine-tuning ChatGPT for automatic scoring. *Computers and Education: Artificial Intelligence*, 6, 100210.
- Lee, G. G., Latif, E., Wu, X., Liu, N., & Zhai, X. (2024). Applying large language models and chain-of-thought for automatic scoring. *Computers and Education: Artificial Intelligence*, 6, 100213.
- Levenshtein, V. I. (1966, February). Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics-doklady* (Vol. 10, No. 8).
- Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.
- Mansour, W. A., Albatarni, S., Eltanbouly, S., & Elsayed, T. (2024, May). Can large language models automatically score proficiency of written essays?. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)* (pp. 2777-2786).
- McCaffrey, D. F., Casabianca, J. M., Ricker-Pedley, K. L., Lawless, R. R., & Wendler, C. (2022). Best practices for constructed-response scoring. *ETS Research Report Series*, 2022(1), 1-58.
- Naismith, B., Cardwell, R., LaFlair, G., Nydick, S., & Kostromitina, M. (2025). *Duolingo English Test: Technical manual*. Duolingo Research Report]. Duolingo. <https://go.duolingo.com/dettechnicalmanual>.
- Odell, M. K., & Russell, R. C. (1918). Method of indexing (U.S. Patent No. 1,261,167). U.S. Patent and Trademark Office. <https://patents.google.com/patent/US1261167A>
- Okubo, T., Houlden, W., Montuoro, P., Reinertsen, N., Tse, C. S., & Bastianic, T. (2023). AI scoring for international large-scale assessments using a deep learning model and multilingual data.
- Shin, H. J., Andersen, N., Horbach, A., Kim, E., Baik, J., & Zehner, F. (2024). *Operational automatic scoring of text responses in 2016 ePIRLS: Performance and linguistic variance*.
- Thomas, L. D., Romasanta, A. K. G., & Priego, L. P. (2026). Jagged competencies: Measuring the reliability of generative AI in academic research. *Journal of Business Research*, 203, 115804.
- Zehner, F., Shin, H. J., Kerzabi, E., Horbach, A., Gombert, S., Goldhammer, F., Zesch, T., & Andersen, N. (2025). *Down the cascades of Omethi: Hierarchical automatic scoring in large-scale assessments*. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA)*.

Appendices

A. Standardized Scoring Template

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

Steps

1. Translation:
 - Translate each non-English response into English. Make sure to retain the meaning of original sentence.
 - Translate the complete non-English sentence or phrase to ensure the full meaning is retained.
 - Correct grammar and spelling of English responses.
2. Translation Validation:
 - Verify translation accuracy to avoid "hallucination" (linguistically correct but misrepresenting the intended meaning).
 - If a response is untranslatable, retain the original text.
3. Scoring: Score the response based on the given scoring guide.
4. Scoring Validation: Confirm that the score aligns with the scoring guide criteria.
5. Output:
 - Output the final score, followed by the text in JSON format (English translation or original if untranslatable).
 - Do not provide a rationale.

Output Format:

- 1: {English translation or original text}

0: {English translation or original text}

Passage: {**question-specific passage summary**}

Question: {**target item question**}

Scoring Guide:

Evaluate responses based on the following criteria:

{**optimized scoring guide**}