Down the Cascades of Omethi: Hierarchical Automatic Scoring in Large-Scale Assessments

Fabian Zehner[1,2], Hyo Jeong Shin[3], Emily Kerzabi[4], Andrea Horbach[5], Sebastian Gombert[1], Frank Goldhammer[1,2], Torsten Zesch[6], Nico Andersen[1]
[1]DIPF, Germany, {f.zehner,n.andersen}@dipf.de
[2]Centre for International Student Assessment (ZIB), Germany
[3]Sogang University, South Korea; [4]Educational Testing Service, USA
[5]IPN, Germany; [6]FernUniversität in Hagen, Germany

Abstract

We present the framework Omethi, which is aimed at scoring short text responses in a semi-automatic fashion, particularly fit to international large-scale assessments. We evaluate its effectiveness for the massively multilingual PISA tests. Responses are passed through a conditional flow of hierarchically combined scoring components to assign a score. Once a score is assigned, hierarchically lower components are discarded. Models implemented in this study ranged from lexical matching of normalized texts—with excellent accuracy but weak generalizability—to fine-tuned large language models—with lower accuracy but high generalizability. If not scored by any automatic component, responses are passed on to manual scoring. The paper is the first to provide an evaluation of automatic scoring on multilingual PISA data in eleven languages (including Arabic, Finnish, Hebrew, and Kazakh) from three domains ($n = 3.8$ million responses). On average, results show a manual effort reduction of 71 percent alongside an agreement of $\kappa = .957$, when including manual scoring, and $\kappa = .804$ for only the automatically scored responses. The evaluation underscores the framework's effective adaptivity and operational feasibility with its shares of used components varying substantially across domains and languages while maintaining homogeneously high accuracy.

1 Introduction

A river adapts its flow to diverse exterior conditions, by meandering, or alternating its velocity and depth, to reach its target inevitably and naturally. In this paper, we propose the hierarchical, response-adaptive framework *Omethi* for automatically scoring short text responses from assessments. The proposed framework is named after the Omethi River, for it is similarly responsive by combining modern and baseline scoring methodology adaptively at the response level, while contending with diverse languages and multiple assessment domains in an operational setting and distinct quality requirements. Large-scale assessments, especially international ones (e.g., PISA, the *Programme for International Student Assessment*; OECD, 2023), pose diverse conditions to automatic scoring (Zesch et al., 2023), similar to the varied surroundings a river is exposed to. In turn, automatic scoring encompasses a range of approaches with particular strengths and weaknesses (see Galhardi and Brancher, 2018; Gao et al., 2024).

Accordingly, the paper provides three major contributions. First, we present a novel hierarchical composition of models for automatically scoring short text responses, particularly fit to the complex settings present in large-scale assessments.

Second, for a first implementation of the framework, we propose a hierarchical collection of models, including a new rigorous method with weak generalizability, called *Fuzzy Lexical Matching* (FLM), alongside fine-tuned XLM-RoBERTa (XLM-R; Conneau et al., 2020) and support vector machine classifiers (SVM; Cortes and Vapnik, 1995). Human raters, integral to assessment operations, serve as the final component in the sequence of scoring methods presented here, turning the implemented pipeline into a semi-automatic system.

Third, this is the first paper to evaluate automatic scoring on massively multilingual data from PISA tests including all three major domains (i.e., *reading*, *mathematics*, and *science*; OECD, 2024). With the complete dataset containing 59 test languages from 86 countries and regions in total, we sampled a subset of 11 test languages for the present evaluation, resulting in about 3.8 million text responses to 160 items from 3 assessment domains and more than 270,000 students. To represent diverse language families and writing systems, the selected test languages included Arabic, Finnish, Hebrew, Kazakh, and Korean, among others.

The empirical evaluation was guided by two

Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications, pages 660–671

overarching research questions. (I) Overall and for each subcomponent, how effective is the model at generating accurate scores and reducing manual effort? (II) How robust are scoring accuracy and reduced manual effort across subsamples with different test languages?

2 Background

2.1 Relevance for Operational Assessments

International educational large-scale assessments, such as PISA, are characterized by their large scope in addressing diverse student characteristics from different cultures using complex methodology. This can pose significant challenges for automatic scoring (Yan et al., 2020; Zesch et al., 2023). The resulting diversity manifests in response texts and corresponding scoring, stemming from many factors, including the world-wide participation (i.e., over ninety countries and economies in PISA 2025; OECD, 2025). The tests are administered in a large number of test languages (almost sixty test languages from 2018 to 2022), with high-resource languages, such as Indonesian, just as low-resource languages, such as Kazakh or Catalan. Moreover, the tests assess three major literacy domains, using a large number of items and various item types with complex coding guides for constructed-response formats. Additionally, the low-stakes nature at the individual level often results in lower test engagement (Schlosser et al., 2019) and, thus, more informal, fragmented, and less integrated (Chafe, 1982) text responses. Continuous changes in assessment design—such as the transition from paper- to computer-based testing and the adoption of adaptive testing—introduce additional variability over time; for example, by reducing the number of responses per item (OECD, 2024) or by impacting the length and quality of text responses (Zehner et al., 2019, 2020). On top of this, not only sample sizes vary largely per test language (e.g., from $n = 269$ to $n = 22{,}163$ responses per item in the present paper's reported dataset), which poses challenges for training, but also a reduced rigor in human coding can lead to more label noise in subsamples. At the same time, large-scale assessments pose incontestable quality requirements (see OECD, 2025), including high-quality coding and accountability (i.e., explainability), due to their high stakes at the state level. Shin et al. (2019) demonstrated that automatic scoring can align closely with human experts in identifying rater severity, and less so regarding centrality and accuracy, highlighting further challenges in introducing automatic systems in operational procedures. Noteworthy, large-scale assessments usually administer a subset of items repeatedly over time, making them an attractive field of application for supervised learning.

Thus far, automatic scoring has seen limited research and operational use in international large-scale assessments. Early efforts include the introduction of PISA'a Machine-Supported Coding System (Yamamoto et al., 2018), a precursor to FLM, and a baseline evaluation for German (Zehner et al., 2016). Recent research funded by international bodies, such as on IEA's ePIRLS data (*International Association for the Evaluation of Educational Achievement*; Shin et al., 2024), and a competition on data from the National NAEP (*National Assessment of Educational Progress*; Whitmer et al., 2023) signal growing interest in automating scoring, notoriously centering around national U.S. assessments (Yan et al., 2020).

2.2 Diverse Models to Address Text Diversity

All these extraneous factors manifest in varying degrees of linguistic variance in text responses (Zesch et al., 2023; Horbach and Zesch, 2019) across cohorts, subpopulations (i.a., languages), domains, items, and their context. Single automatic scoring approaches can fall short of adequately addressing this diversity. For instance, while lexical matching methods offer excellent accuracy for known responses, they lack generalizability to unseen linguistic expressions. Moreover, supervised classifiers are often hampered as they assign a label regardless of relatively low probabilities (i.e., confidence) for certain instances (Li et al., 2023).

Recognizing these limitations, the here presented first collection of implemented components in an Omethi framework retain human raters as the final recourse when automatic models fail to score responses with sufficient confidence, rendering it a semi-automatic system. By hierarchically composing multiple scoring approaches and discarding lower-level components once a score is confidently assigned, Omethi navigates the complexities of international large-scale assessments while maintaining the high-quality standards required for them.

2.3 Ensembles for (Semi-)Automatic Scoring

Ensembles for automatic and semi-automatic scoring come in two fashions: algorithmic ensembles that inherently comprise multiple models (e.g., ran-

dom forests) or combinations of relatively loosely coupled models (e.g., stacking). Omethi belongs to the latter and diverges from traditional systems by combining multiple components, including supervised classifiers, in a conceptually governed, top-down manner rather than relying on data-driven, bottom-up learning. Unlike the common paradigm of identifying a single optimal model for a dataset, task, or domain, Omethi deliberately alternates models at the response level based on explicit criteria. This approach contrasts with standard ensembles, such as those in Goenka et al. (2020) and Ormerod (2022), where model selection is carried out uniformly (e.g., majority voting or averaging) or with ensembles designed to capture diverse response characteristics (e.g., Mohler et al., 2011; Sahu and Bhowmick, 2020; Sakaguchi et al., 2015; Zhang et al., 2022). For instance, Heilman and Madnani (2013) stacked models for item-specific n-gram features and text similarity, while Roy et al. (2016) employed transfer learning between general- and question-specific classifiers.

If humans are still involved during inference, the scoring is considered semi-automatic. For systems deferring responses to humans, appropriate confidence thresholds of the automatic component need to be identified; referred to as *deferral policy* in (Li et al., 2023), which we rephrase here as the *eligibility policy* for assigning a score. This has been investigated for semi-automatic systems, such as in Andersen et al. (2023) and Horbach et al. (2014), which combined unsupervised clustering with human scoring. Horbach and Pinkal (2018) more directly integrated humans and machines via semi-supervised clustering. In the context of label probabilities as a confidence criterion, results on identifying optimal confidence thresholds have been mixed. Suen et al. (2023) successfully set thresholds based on a minimum required F_1 score, while Bexte et al. (2024) observed substantial item- and data-wise variation in confidence distributions with this, failing to identify viable thresholds for certain items at all. Funayama et al. (2022) similarly used confidence scores to revert to human raters, and Li et al. (2025) proposed a constant threshold of $\delta = .25$, basically halving the range of values from random chance to perfect agreement.

3 Omethi Framework

Unlike traditional ensemble methods, Omethi orchestrates scoring components hierarchically based on their conceptual priority, compiling a logical decision flow that is informed by each component's inherent characteristics. If a component is eligible by satisfying its specific conditions (i.e., its eligibility policy), it assigns the final score and the response bypasses subsequent components.

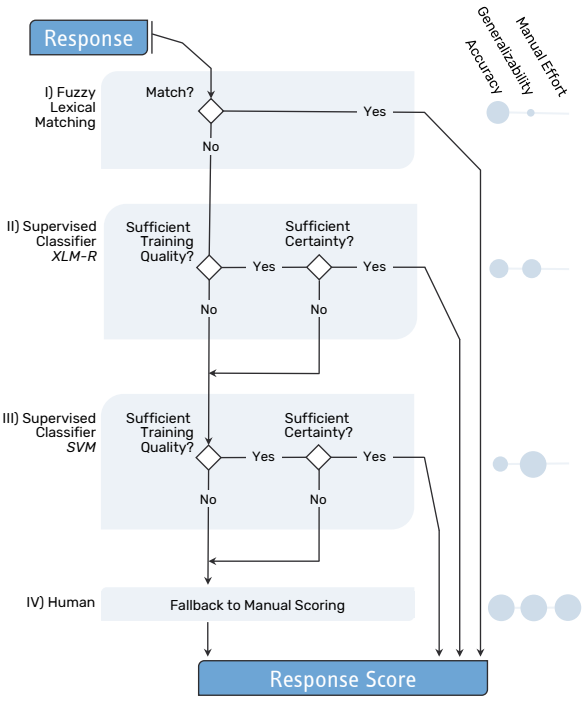

Figure 1: Response flow through the implemented Omethi pipeline

In this paper, we present an initial implementation consisting of four components, described in the following (see Figure 1). The rationale underlying the implementation was to allocate components with the highest accuracy prior to those exhibiting higher generalizability, while also aiming at minimizing human effort resulting from responses deferred to the final component.

During training, each scoring component was built separately using data from a given subsample; that is, for (i) a specific item and (ii) test language. Human scores available from the operational PISA studies served as the ground truth for training. During inference, each response was first evaluated by Fuzzy Lexical Matching (FLM), which attempts to match normalized text to a pool of normalized response texts. If sufficient matching responses were identified that satisfied predefined score-homogeneity criteria (see next section), propagating their score to the unseen response is considered highly reliable, as the response had been

scored multiple times previously by humans, or at least a lexically very close counterpart. FLM, therefore, receives the highest priority in the flow because its classifications are largely valid, interpretable, and applicable for any language. However, FLM's obvious downside is its lack of out-of-sample generalization, the severity of which depends on the item-specific linguistic variance in the responses. Thus, if FLM *could* provided a score, that score *was* adopted, bypassing subsequent components. Otherwise, the response proceeded to the next scoring component.

Responses not scored by FLM were next passed to a supervised classifier: a fine-tuned XLM-RoBERTa classifier or SVM. The model's output was assigned to the response if its overall training performance quality sufficed and the individual classification's confidence exceeded an item- and language-specific threshold.

If none of the automatic components satisfied their eligibility policy, the response was forwarded to the final component, namely manual scoring by human raters.

3.1 Fuzzy Lexical Matching

FLM extends the idea of PISA's Machine Support Coding System (Yamamoto et al., 2018), operationally introduced in PISA 2018. There, strict exact string matching was applied, automatically propagating scores if at least five homogeneously scored text responses were found in legacy data.

FLM builds on this widely adoptable principle of matching unseen to historic data. In contrast to exact matching, FLM first normalizes the texts by traditional preprocessing techniques. The normalization pipeline was first evaluated on ePIRLS data (the *Progress in International Reading Literacy Study*; Shin et al., 2024). The standard techniques used were white-space trimming, punctuation removal, case insensitivity, diacritics removal, stemming, stop word removal, and bag of words.

For optimization to a subsample (i.e., item and language), this set of normalization techniques is trained on the respective data. That is, the effectiveness of each pipeline step is evaluated using the coefficient ER (Effort Reduction), simply constituting the share of matched responses, $ER = \frac{n_m}{n_t}$; n_t denoting the total number of responses in the data and n_m the number of matches. Importantly though, FLM's scoring quality also manifests in ER because the method requires sufficiently frequent as well as homogeneously scored responses for automatic scoring. That is, if the grouping of the normalized texts leads to heterogeneous scores within that group, ER will decrease. A response is automatically scored if the following criteria are met. For a given response i, let m_i denote the number of its matches and s_i the number of responses that received the dominant score in the group. Then, the response is scored ($M = 1$) or not scored automatically ($M = 0$) as follows:

$$M = \begin{cases} 1, & \text{if } m_i \geq 3 \text{ and} \\ & \quad s_i \geq \max\left(\lceil m_i \cdot .92 \rceil, m_i - 5\right) \\ 0, & \text{otherwise} \end{cases}$$

That is, a response is scored automatically if at least 3 responses are matched, requiring a minimum of 92 percent of homogeneous scores, but limited to an absolute maximum of 5 deviant responses.[1]

Whenever a pipeline step in FLM leads to a decrease in ER, the respective step is discarded for the specific subsample (i.e., item and language). For example, if respondents were asked to provide an email address from a text, applying punctuation removal on the responses eliminates relevant information, leading to heterogeneously scored matching groups, a reduced ER, and, hence, this normalization step would be discarded during inference.

Another adaptive step in FLM is the tailoring of stopword lists to the subsample. The rationale behind this is twofold. For one, stopword lists are language-specific and differ largely in their scope. Second, whether certain words are predictive for a response's score depends on the item. Therefore, if an optimized stopword list leads to an increase in ER or an increase of the overall accuracy while ER remains identical, the optimized stopword list is used during inference.

3.2 Supervised Classifiers

Two types of classifiers based on supervised learning were built: fine-tuned XLM-RoBERTa models and support vector machines. During inference, both only take response texts as their input, not considering item stems, stimulus materials, or scoring guides.

As a core component, fine-tuned XLM-RoBERTa models (Conneau et al., 2020) were employed for their robust multilingual representation and classification capabilities. XLM-R is a massively multilingual model pretrained on a corpus

[1] In PISA, the minimum inter-rater agreement is required to be 92 percent (OECD, 2024).

comprising one hundred languages. For enabling binary (i.e., dichotomous) and multiclass (i.e., polytomous) scoring, respectively, a classification head was appended to the pretrained model.

With the objective to only have the model assign fairly probable scores, labels' output probabilities were stored for each instance. Using Receiver Operating Characteristic (ROC) analysis, an optimal threshold of label output probability o_j, specific to subsample j, was determined to minimize misclassifications. This threshold was determined by maximizing Youden's index (Youden, 1950), which quantifies the trade-off between sensitivity and specificity. Specifically, we computed

$$o_j = \arg\max_{x \in [0,1]} \left(\frac{\mathrm{TP}_x}{\mathrm{TP}_x + \mathrm{FN}_x} - \frac{\mathrm{FP}_x}{\mathrm{FP}_x + \mathrm{TN}_x} \right),$$

where TP, FP, TN, and FN denote subsample j's number of true positives, false negatives, and so on, based on a vector of classification correctness at threshold x.

This threshold identification differs from conventional ROC analyses, which typically rely on the actual binary labels rather than their correctness. With tailored confidence thresholds, the XLM-R classifiers ensure reliable predictions while deferring uncertain cases to downstream components. Moreover, only classifiers with sufficient training performance were employed at all.

In addition to fine-tuned XLM-R classifiers, support vector machine classifiers were trained using XLM-R embeddings as the input features. With a small number of entirely underfitting XLM-R models, the SVM classifiers were designed as fallback classifiers before ultimately deferring to human scoring. While linguistic representation remained consistent with XLM-R classifiers, SVMs' distinct classification provided—despite somewhat poorer accuracy—more robustness in scenarios where datasets may be small, noisy, or skewed in their class distribution.

As the threshold for inference certainty, SVM classifiers used the arithmetic mean probability instead of the ROC-based approach employed for fine-tuned XLM-R models. This simpler thresholding mechanism was chosen because SVMs were applied only to responses that had already been deemed uncertain by upstream models.

4 Empirical Evaluation

Omethi implemented as described above was evaluated by simulating its flow on a real-world dataset.

4.1 Dataset and Instrument

In PISA (OECD, 2023), 15-year-old students take tests in a total of three domains to assess their scientific, mathematics, and reading literacy. For the present study, we had available text responses for all construct-response items from all Field Trials and Main Studies for PISA 2018 and 2022. With the complete data being too large for one evaluation and its reporting, we sampled 11 datasets with diverse languages for the present paper: Arabic (Jordan), Traditional Chinese (Chinese Taipei), Finnish (Finland), English (U.S.), German (Germany), Hebrew (Israel), Indonesian (Indonesia), Kazakh (Cyrillic script; Kazakhstan), Korean (South Korea), Portuguese (Brazil), and Spanish (Spain). Corresponding to $n = 270{,}445$ students, this resulted in a total of $n = 3{,}773{,}728$ responses that had already been assigned human scores in PISA with its high quality standards (OECD, 2025).

The dataset comprised 160 items (89 reading, 39 math, and 32 science items), 121 with two and 39 of them with three score levels. Not all items had been administered in all selected languages, resulting in a total of 1,676 datasets (i.e., classifiers to be trained). Sample items with corresponding coding guides can be found on the OECD's website (OECD, 2025). Coding guides for some items are simple, such as *"Full credit is given when the student states that the weight or size [. . .] was not provided . . ."* (CR548Q09), others are more complex, such as *"Selects one of the names and gives an appropriate explanation as described below."* (with 19 explanations specified and mapped to one of three different names; CR557Q14).

Table 1 shows exemplary responses for each domain. They are selected from coding guides released by the OECD and not from the evaluation data set, because items in PISA are confidential due to the assessment's high stakes at the national level, constraining the selection options. Note that constructed-response items in math regularly involve mathematical reasoning (sometimes, naming a number), but rarely involve stating formulas.

4.2 Implementation

We used Python 3.11.5 and R 4.4.3 (R Core Team, 2025). For XLM-R, the base model[2] with 279 million parameters was used. Due to the large number of required classifiers, hyperparameters and

[2] https://huggingface.co/FacebookAI/xlm-roberta-base [2025-04-01]

Domain	Item ID	Item Stem	Sample Response	Context
Math	CMA159Q01	Peter thinks there is a greater probability of the arrow stopping on blue in Spinner A than there is in Spinner B. Is Peter correct?	Because $\frac{1}{2} = \frac{2}{4}$. He is not correct because the probability is the same for each spinner.	Details (OECD, 2025)
Reading	CR548Q09	With whom do you agree?	Sam. These are only two texts and more research is needed before a conclusion can be made.	Details (OECD, 2025)
Science	CS623Q03	What is the biological reason for this effect?	Increasing sweat levels in high temperatures keeps the body from getting too hot.	Details (OECD, 2024)

Table 1: Sample PISA items and responses from released coding guides

settings were not tuned classifier-wise for computational constraints. Instead, a fixed batch size of $B = 32$, learning rate of $\eta = 5e^{-5}$, the AdamW optimizer (Loshchilov and Hutter, 2017), and a cosine learning rate scheduler with warm-up were used. Training was capped at ten epochs, with early stopping after stagnating performance in three consecutive epochs. Cross-entropy loss was used for optimization. SVMs employed a Radial Basis Function (RBF) kernel.

Classifiers were deployed only if they met minimum training performance thresholds ($\kappa \geq .300$ for XLM-R and $\kappa \geq .900$ for SVM) as measured by QWK (Quadratic Weighted Kappa; Cohen, 1960). F_1-score is reported as F_1 micro. Importantly, in the reported evaluation, the final manual scoring component was assumed to yield perfect accuracy, despite normal inconsistencies in human scoring. That is, this component takes the ground truth label, as provided by PISA's human raters, as its output. This assumption was made for two reasons: (i) consistent estimates of inter-rater agreements were not available for all subsamples, and (ii) the substantial reduction in manual effort could alter relevant rater cognition (Bejar, 2017) and reliability (Padó and Padó, 2022).

Finally, due to computational constraints stemming from the fine-tuning of many XLM-R models, an 80/20 training-test-split was used for evaluation.

4.3 Results

All reported average values constitute means weighted by sample size across all classifiers.[3]

[3]For the sake of readability, only a selected set of standard errors is reported (in brackets) where comparisons may be relevant. All result data is available upon request.

	Acc (%)	κ	F_1	*ER* (%)
Math	98.8	.972	.988	73.0
Reading	97.7	.954	.977	71.0
Science	97.7	.951	.977	67.3

Table 2: Omethi's performance by domain

4.3.1 Performance by Domain

Omethi achieved very high agreement with human scores, with an average $\kappa = .957$, $TPR = .968$, and $FPR = .977$, alongside substantial manual effort savings (on average, $ER = 70.5\%$). Notably, these results include a share of responses scored manually, as reflected in the effort reduction metric, and assume perfect agreement for this subset. Nonetheless, the reported figures represent the expected scoring quality if Omethi were deployed in an operational setting.

Table 2 details the agreement and effort reduction across all domains. Scores showed the highest agreement for math items with an average accuracy of 98.8 percent [$\pm 0.1\%$], $\kappa = .972$ [$\pm.003$], and an effort reduction of 73.0 percent [$\pm 1.2\%$], meaning 27.0 percent of responses have been deferred to human scoring. For the other two assessment domains, Omethi scores showed marginally lower but still very high agreement values, with accuracy at 97.7 percent [$\pm 0.1\%$] for reading and identical 97.7 percent [$\pm 0.3\%$] for science ($\kappa = .954$ [$\pm.002$] and $\kappa = .951$[$\pm.005$]), and an effort reduction of 71.0 percent [$\pm 0.8\%$] and, for science somewhat lower, 67.3 percent [$\pm 1.1\%$], respectively.

The overall high agreement for the majority of classifiers across domains and languages with only rare exceptions is visualized in Figure 2.

Distinguishing performance of individual automatic components, Table 3 reports component-

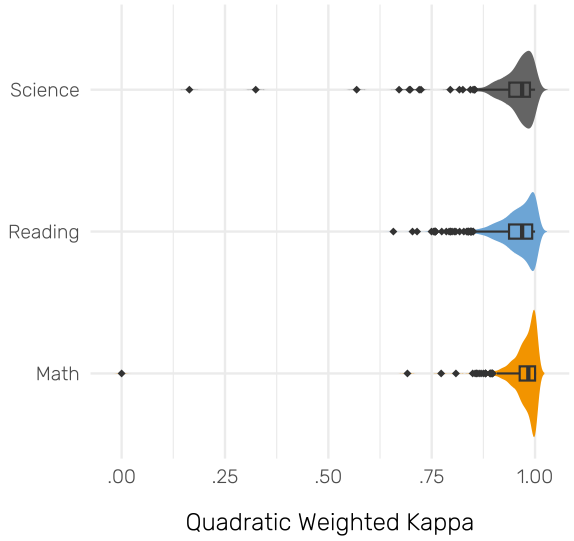

Figure 2: Omethi's distribution of performance at the item- and language-level across domains

wise agreement values, excluding manual scoring during inference. The first three columns show the agreement of the components' automatic score with human scores for the subset of responses that were scored by the respective component for Omethi. For FLM, accuracy was at $Acc_{FLM} = 98.6$ percent on average, XLM-R's performance was $\kappa_{XLM} = .775$ [$\pm.011$], and for SVM, $\kappa_{SVM} = .554$. The fourth column (κ_{auto}) shows agreement for all automatic components combined, again only for the subset of responses scored automatically; $\kappa_{auto} = .804$ [$\pm.009$]. The last column ($\kappa_{\text{XLM}_{\text{all}}}$) reports agreement as would have been the case if all responses were scored by XLM-R classifiers alone, showcasing the substantial added value of the combination of automatic scoring approaches beyond transformer finetuning as displayed in the adjacent column on the left ($\Delta\kappa_{auto,XLM_{all}} = .093$). The Appendix visualizes this gain of the Omethi pipeline over mere XLM-R fine-tuning (see Figure 4). Similarly, Figure 5 (see Appendix again) shows that XLM-R outperforms SVM for the majority of item- and language-specific classifiers, which is in line with the rationale underlying the component hierarchy, while also showcasing the number of random-level XLM-R classifiers, for which SVM was added as a potential fallback.

4.3.2 Robustness across Languages

Table 4 displays Omethi's agreement with human scores and effort reduction across subsamples,

	Acc_{FLM}	κ_{XLM}	κ_{SVM}	κ_{auto}	$\kappa_{\text{XLM}_{\text{all}}}$
Math	99.2	.715	.414	.765	.683
Reading	98.2	.792	.584	.845	.739
Science	98.8	.784	.600	.756	.677

Table 3: Performance for automatic components and their combination (*auto*); responses deferred to manual scoring excluded, except for XLM_{all}

which includes, among others, different test languages. Overall agreement was homogeneously high, with accuracy ranging mainly from 97.8 (Spain) to 99.0 percent (Jordan) and the exception of Indonesia with 96.9 percent. In contrast, effort reduction varies largely from 60.3 (Indonesia) to 76.1 percent (Chinese Taipei), showing how the implemented scoring conditions effectively identified instances that required human scoring. Similarly, the shares of responses scored by different components varied heterogeneously across subsamples (see Appendix A, Table 6). For example, FLM scored 29.5 percent of responses in the subsample from Chinese Taipei (Traditional Chinese), which stands in stark contrast to the one from Israel (Hebrew) with only 19.7 percent. For Spain (Spanish), SVMs only scored 1.2 percent of the responses, compared to Jordan (Arabic) with 9.8 percent combined with an outlier of only 28.1 percent of sufficiently confident scoring by XLM-R, whereas the XLM-R classifiers for the U.S. (English) scored even 52.2 percent of the responses.

	Acc (%)	κ	F_1	*ER* (%)
ara-jor	99.0	.965	.990	66.7
deu-deu	98.1	.961	.981	72.2
eng-usa	98.0	.959	.980	75.5
esp-esp	97.8	.957	.978	70.4
fin-fin	98.5	.969	.985	75.2
heb-isr	98.1	.961	.981	70.8
ind-idn	96.9	.928	.969	60.3
kaz-kaz	98.3	.962	.983	67.9
kor-kor	98.3	.965	.983	75.1
por-bra	98.4	.965	.984	72.0
zho-tap	98.1	.963	.981	76.1

Table 4: Performance and effort reduction by language, incl. manual scoring

4.3.3 Component Shares

Table 5 shows the percentage of responses scored by the respective component due to meeting the eligibility policy. FLM and XLM-R played the major role for automatic scoring. With its position at the end of the sequence of automatic components, SVMs only played a minor role quantitatively. Nevertheless, as shown in Table 6, there were settings, such as the subsample from Jordan (Arabic) in which the first automatic components do not perform well and SVM takes over some of the shares to retain the homogeneously high level of accuracy.

The prevalence of different flows responses take through the scoring components is displayed in Figure 3. Said cases where XLM-R and FLM do not manage to score responses for which SVM takes over are visible in the figure as the orange ribbon. Moreover the figure disentangles the specific conditions for why responses are not scored by specific components.

5 Discussion

The results demonstrate Omethi's effectiveness in orchestrating multiple methods in an explicitly designed, adaptive scheme for automatic scoring across domains and languages while maintaining uniformly high accuracy. With an average agreement of $\kappa = .957$ compared to complete human scoring and manual effort reductions of 70.5 percent across domains, Omethi proves its feasibility in and operational usefulness. Thus far, for PISA data, effort reduction gains have been reported to be smaller with other methods and data sets comprising fewer test languages and assessment domains (Andersen et al., 2023; Yamamoto et al., 2018). Critically, the system's hierarchical composition and scoring conditions ensured that accuracy was prioritized, resulting in varying effort reduction across settings. It is important to note that *manual effort reduction* here does not refer to the entirety of human involvement in operational assessment procedures but only the share of automatically scored responses during inference.

	FLM	XLM-R	SVM	Manual
Math	35.0	34.2	3.8	27.0
Reading	24.1	44.5	2.5	29.0
Science	18.4	46.0	2.8	32.7

Table 5: Proportions (%) of component usage

The importance of combining different methodologies was evidenced by the homogeneous accuracy levels despite heterogeneous shares of responses being scored by different components across domains and languages. Each component in Omethi played a distinct role, contributing to the system's overall robustness. While FLM and XLM-R dominated the scoring, partly due to their position in the sequence, SVMs served as a crucial fallback mechanism, stepping in when upstream components failed to score confidently. Although SVMs scored only a minor share of responses quantitatively, their role turned out as indispensable in maintaining accuracy for certain subsamples. This underscores the importance of the adaptive workflow, where eligibility policies diagnosed the risk of misclassifications and led to passing on responses.

For identifying a confidence threshold as components' eligibility policy, the proposed maximizing of the ROC-based Youden's Index on misclassifications worked excellently for XLM-R classifiers. Less so for SVM classifiers that were faced with only the more challenging responses not scorable by upstream components. Hence, this measure may be added to the repertoire of threshold identification methods, complementing fixed constants (e.g., Li et al., 2025) or the definition of minimum F_1 scores as proxies (e.g., Bexte et al., 2024), but its suitability needs to be verified.

Omethi's strength in adaptivity may also introduce challenges in ensuring equivalence and fairness across test languages and subpopulations, necessitating careful validation and bias checks, as bias is known to be potentially masked at the aggregate level (Andersen et al., 2023). From an operational standpoint, implementing Omethi in international large-scale assessments would require a rigorous quality monitoring.

For human raters, the implementation of such a framework would result in multiple changes that may affect rater cognition in positive or negative ways, or both. First, the number of responses decreases, potentially leading to less fatigue, monotonous work, and slippage. Second, raters' oversight of frequent responses would diminish and would thus change so-called contrast or context effects by preceding responses, an effect repeatedly found even in highly standardized settings with well-trained raters (Attali, 2011; Meadows and Billington, 2005). Third, both for automatic systems as well as humans (Padó and Padó, 2022), incorrect responses are more challenging to score.

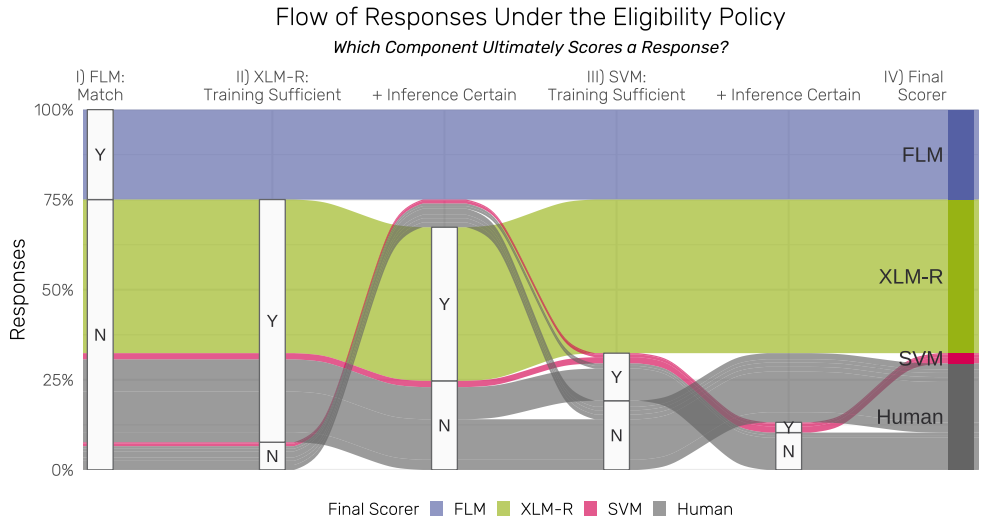

Figure 3: Share of response flows through Omethi's components. Y/N (yes/no) = eligibility policy satisfied?

Accordingly, if classifiers with poor recall leave human raters with a higher frequency of incorrect responses, this also may show effects.

In conclusion, Omethi's response-level adaptivity, combined with its capability to maintain high accuracy across diverse contexts, positions it as a powerful tool for operational employment in assessments. While challenges may remain in, among others, ensuring fairness, the system's effectiveness and flexibility pave the way for its operational use and future enhancements. In follow-up studies, the results presented here may be used to sample specific datasets informative to diverse facets to carry out an evaluation in order to systematize performance differences and identify optimal hyperparameters, respectively.

Ethical Considerations

The implementation of a framework such as Omethi in large-scale assessments necessitates careful attention to ethical principles, which is not always at the forefront of attention (Holmes et al., 2022). Fairness and the mitigation of bias are paramount, as variability in component usage across languages and cultures could lead to disproportionate disadvantages for certain groups. Rigorous validation and bias investigations are essential to ensure equitable performance across diverse populations. Transparency in the scoring process is critical to fostering trust among stakeholders, including organizations such as the OECD, policymakers, and test takers. Clear documentation of the system's decision-making mechanisms and limitations must be provided to ensure interpretability and accountability. Additionally, equity in resource allocation must be discussed, as disparities in system performance between high- and low-resource languages could exacerbate existing inequalities. Finally, the increasing automation of standardized assessments raises broader questions about their role in education. While automation enhances efficiency and scalability, it also risks amplifying uniformity, potentially overlooking diversity facets.

Limitations

The study faces several limitations, primarily due to computational constraints. With many test languages, items, and domains, a large number of item- and language-specific classifiers were fine-tuned using the XLM-RoBERTa base model. This scale rendered classifier-specific hyperparameter tuning via grid search computationally infeasible, necessitating the use of fixed hyperparameters. Similarly, k-fold cross-validation was not conducted due to resource limitations, restricting the evaluation to a single 80/20 train-test split.

The system's runtime scales with the number of components, complicating potential real-time deployment in certain settings. Additionally, the evaluation focused exclusively on operational data, lacking comparison with public benchmarks or standardized datasets. The use of human scoring as the gold standard, particularly for responses deferred to manual scoring, assumes perfect inter-rater reliability, which may overestimate accuracy in production.

References

Nico Andersen, Fabian Zehner, and Frank Goldhammer. 2023. Semi-automatic coding of open-ended text responses in large-scale assessments. *Journal of Computer Assisted Learning*, 39(3):841–854.

Yigal Attali. 2011. Sequential effects in essay ratings. *Educational and Psychological Measurement*, 71(1):68–79.

Isaac I. Bejar. 2017. A historical survey of research regarding constructed-response formats. In Randy Elliot Bennett and Matthias von Davier, editors, *Advancing Human Assessment*, pages 565–633. Springer, Cham.

Marie Bexte, Andrea Horbach, Lena Schützler, Oliver Christ, and Torsten Zesch. 2024. Scoring with confidence? – Exploring high-confidence scoring for saving manual grading effort. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 119–124, Mexico City, Mexico. Association for Computational Linguistics.

W. L. Chafe. 1982. Integration and involvement in speaking, writing, and oral literature. *Spoken and Written Language: Exploring Orality and Literacy*, pages 35–54.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451. Association for Computational Linguistics, Stroudsburg, PA, USA.

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine Learning*, 20(3):273–297.

Hiroaki Funayama, Tasuku Sato, Yuichiroh Matsubayashi, Tomoya Mizumoto, Jun Suzuki, and Kentaro Inui. 2022. Balancing cost and quality: An exploration of human-in-the-loop frameworks for automated short answer scoring. In Maria Mercedes Rodrigo, Noburu Matsuda, Alexandra I. Cristea, and Vania Dimitrova, editors, *Artificial Intelligence in Education*, volume 13355, pages 465–476. Springer International Publishing, Cham.

Lucas Busatta Galhardi and Jacques Duílio Brancher. 2018. Machine learning approach for automatic short answer grading: A systematic review. In Guillermo R. Simari, Eduardo Fermé, Flabio Gutiérrez Segura, and José Antonio Rodríguez Melquiades, editors, *Advances in Artificial Intelligence – IBERAMIA 2018*, volume 11238, pages 380–391. Springer International Publishing, Cham. Series Title: Lecture Notes in Computer Science.

Rujun Gao, Hillary E. Merzdorf, Saira Anwar, M. Cynthia Hipwell, and Arun R. Srinivasa. 2024. Automatic assessment of text-based responses in post-secondary education: A systematic review. *Computers and Education: Artificial Intelligence*, 6(100206).

Palak Goenka, Mehak Piplani, Ramit Sawhney, Puneet Mathur, and Rajiv Ratn Shah. 2020. ESAS: Towards practical and explainable short answer scoring (student abstract). *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(10):13797–13798.

Michael Heilman and Nitin Madnani. 2013. ETS: Domain adaptation and stacking for short answer scoring. In *Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 275–279.

Wayne Holmes, Kaska Porayska-Pomsta, Ken Holstein, Emma Sutherland, Toby Baker, Simon Buckingham Shum, Olga C. Santos, Mercedes T. Rodrigo, Mutlu Cukurova, Ig Ibert Bittencourt, and Kenneth R. Koedinger. 2022. Ethics of AI in education: towards a community-wide framework. *International Journal of Artificial Intelligence in Education*, 32(3):504–526.

Andrea Horbach, Alexis Palmer, and Magdalena Wolska. 2014. Finding a tradeoff between accuracy and rater's workload in grading clustered short answers. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 588–595, Reykjavik, Iceland. European Language Resources Association (ELRA).

Andrea Horbach and Manfred Pinkal. 2018. Semi-supervised clustering for short answer scoring. In *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Andrea Horbach and Torsten Zesch. 2019. The Influence of Variance in Learner Answers on Automatic Content Scoring. *Frontiers in Education*, 4:4.

Yuheng Li, Mladen Raković, Namrata Srivastava, Xinyu Li, Quanlong Guan, Dragan Gašević, and Guanliang Chen. 2025. Can AI support human grading? Examining machine attention and confidence in short answer scoring. *Computers & Education*, 228:105244.

Zhaohui Li, Chengning Zhang, Yumi Jin, Xuesong Cang, Sadhana Puntambekar, and Rebecca J. Passonneau. 2023. Learning when to defer to humans for short answer grading. In *Artificial Intelligence in Education*, pages 414–425, Cham. Springer Nature Switzerland.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Michelle Meadows and Lucy Billington. 2005. A review of the literature on marking reliability.

Michael Mohler, Razvan C. Bunescu, and Rada Mihalcea. 2011. Learning to grade short answer questions using semantic similarity measures and dependency graph alignments. In *Annual meeting of the association for computational linguistics*.

OECD. 2023. *PISA 2022 results (volume I): The state of learning and equity in education*. PISA. OECD.

OECD. 2024. PISA 2015 Released Field Trial Cognitive Items.

OECD. 2025. PISA test. Accessed: 2025-04-15.

Christopher M. Ormerod. 2022. Short-answer scoring with ensembles of pretrained language models. *ArXiv*, abs/2202.11558.

Ulrike Padó and Sebastian Padó. 2022. Determinants of grader agreement: an analysis of multiple short answer corpora. *Language Resources and Evaluation*, 56(2):387–416.

R Core Team. 2025. R: A language and environment for statistical computing. Manual, R Foundation for Statistical Computing, Vienna, Austria.

Shourya Roy, Himanshu S. Bhatt, and Y. Narahari. 2016. An iterative transfer learning based ensemble technique for automatic short answer grading. *ArXiv*, abs/1609.04909.

Archana Sahu and Plaban Kumar Bhowmick. 2020. Feature engineering and ensemble-based approach for improving automatic short-answer grading performance. *IEEE Transactions on Learning Technologies*, 13(1):77–90.

Keisuke Sakaguchi, Michael Heilman, and Nitin Madnani. 2015. Effective feature integration for automated short answer scoring. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1049–1054, Denver, Colorado. Association for Computational Linguistics.

Analia Schlosser, Zvika Neeman, and Yigal Attali. 2019. Differential performance in high versus low stakes tests: Evidence from the GRE test. *The Economic Journal*, 129(623):2916–2948.

Hyo Jeong Shin, Nico Andersen, Andrea Horbach, Eugyum Kim, Jisoo Baik, and Fabian Zehner. 2024. Operational automatic scoring of text responses in 2016 ePIRLS: Performance and linguistic variance.

Hyo Jeong Shin, Edward Wolfe, and Mark Wilson. 2019. Human rater monitoring with automated scoring engines. *Psychological Test and Assessment Modeling*, 61(2):127–148.

King Yiu Suen, Victoria Yaneva, Le An Ha, Janet Mee, Yiyun Zhou, and Polina Harik. 2023. ACTA: short-answer grading in high-stakes medical exams. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 443–447, Toronto, Canada. Association for Computational Linguistics.

John Whitmer, Evelyn Deng, Charles Blankenship, Magdalen Beiting-Parrish, Ting Zhang, and Paul Bailey. 2023. Results of NAEP Reading Item Automated Scoring Data Challenge (Fall 2021). Publisher: EdArXiv.

Kentaro Yamamoto, Qiwei He, Hyo Jeong Shin, and Matthias von Davier. 2018. Development and implementation of a machine-supported coding system for constructed-response items in PISA. *Psychological Test and Assessment Modeling*, 60(2):145–164.

Duanli Yan, André A. Rupp, and Peter W. Foltz, editors. 2020. *Handbook of automated scoring: Theory into practice*. Statistics in the social and behavioral sciences series. CRC Press, Taylor & Francis Group, Boca Raton, FL.

W. J. Youden. 1950. Index for rating diagnostic tests. *Cancer*, 3(1):32–35.

Fabian Zehner, Frank Goldhammer, Emily Lubaway, and Christine Sälzer. 2019. Unattended consequences: How text responses alter alongside PISA's mode change from 2012 to 2015. *Education Inquiry*, 10(1):34–55.

Fabian Zehner, Ulf Kroehne, Carolin Hahnel, and Frank Goldhammer. 2020. PISA reading: Mode effects unveiled in text responses. *Psychological Test and Assessment Modeling*, 62:55–75.

Fabian Zehner, Christine Sälzer, and Frank Goldhammer. 2016. Automatic coding of short text responses via clustering in educational assessment. *Educational and Psychological Measurement*, 76(2):280–303.

Torsten Zesch, Andrea Horbach, and Fabian Zehner. 2023. To Score or not to score: Factors influencing performance and feasibility of automatic content scoring of text responses. *Educational Measurement: Issues and Practice*, 42(1):44–58.

Lishan Zhang, Yuwei Huang, Xi Yang, Shengquan Yu, and Fuzhen Zhuang. 2022. An automatic short-answer grading model for semi-open-ended questions. *Interactive Learning Environments*, 30(1):177–190.

A Appendix

A.1 Proportions of Component Usage by Subsample

	FLM	XLM-R	SVM	Manual
ara-jor	28.8	28.1	9.8	33.3
deu-deu	23.3	46.1	2.9	27.8
eng-usa	20.0	52.2	3.2	24.5
esp-esp	25.8	43.4	1.2	29.6
fin-fin	25.3	46.4	3.5	24.8
heb-isr	19.7	46.8	4.4	29.2
ind-idn	23.0	35.0	2.3	39.7
kaz-kaz	27.5	37.6	2.8	32.1
kor-kor	21.1	50.1	4.0	24.9
por-bra	28.8	39.8	3.4	28.0
zho-tap	29.5	44.1	2.5	23.9

Table 6: Proportions (%) of component usage by subsample

A.2 Gains Beyond Mere XLM-R Fine-Tuning

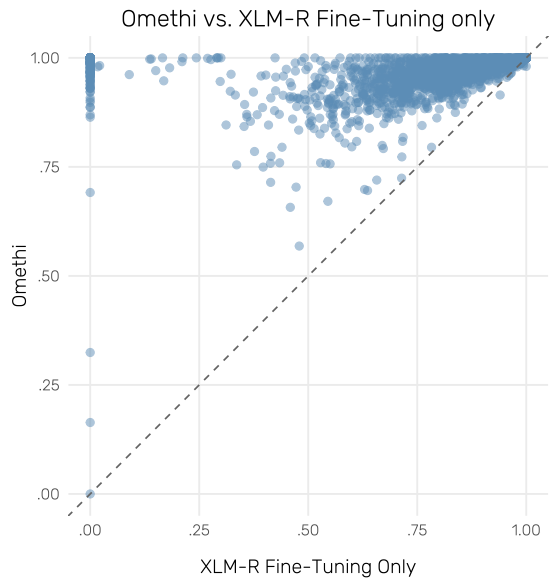

Figure 4: Quadratic Weighted Kappa of XLM-R fine-tuning applied to all responses and Omethi. The complete Omethi pipeline, which includes XLM-R itself and a share of human-scored responses, strongly outperforms XLM-R consistently (values above the diagonal).

A.3 XLM-R Fine-Tuning and SVM

Figure 5: Quadratic Weighted Kappa of XLM-R fine-tuning applied to all responses and SVM. Generally, XLM-R outperforms SVM for the majority of classifiers (values above the diagonal), but SVM shows to be more robust with respect to a number of XLM-R classifiers only showing chance-level performance.

A.4 Acknowledgments

We gratefully acknowledge the OECD for granting access to PISA data, and we thank their PISA Research and Development for Innovation programme, whose funding of a prior project led to the development of the FLM method presented here.