Indian Journal of Information Sources and Services
ISSN: 2231-6094 (P) Vol.16, No.3, 2026, pp.114-121
© The Research Publication, www.trp.org.in
DOI: https://doi.org/10.51983/ijiss-2026.16.3.13

Development of a Natural Language Processing Framework for Automated Assessment of Multilingual Academic Texts

Jabbor Eshonkulov[1*], Ruzikhol Panjieva[2], Ulbolsin Ametova[3], Mastona Yuldasheva[4], Nilufar Yuldosheva[5], Moxira Sadullayeva[6] and Khulkar Kasimova[7]

[1*]Professor, Leading Researcher, Institute of Uzbek Language, Literature and Folklore, Academy of Sciences of the Republic of Uzbekistan, Tashkent, Uzbekistan

[2]Department of Foreign Philology, Termez State University, Termez, Uzbekistan

[3]Associate Professor, Department of Uzbek Literary Studies, Faculty of Journalism and Uzbek Philology, National University of Uzbekistan named after Mirzo Ulugbek, Tashkent, Uzbekistan

[4]Department of Foreign Languages Teaching, Gulistan State Pedagogical Institute, Gulistan, Sirdarya Region, Uzbekistan

[5]Professor, Department of Philology, Karshi State University, Karshi, Uzbekistan

[6]Associate Professor, Department of Uzbek Language and Literature, Karakalpak State University, Nukus, Uzbekistan

[7]Department of Psychology, Mamun University, Khiva, Uzbekistan

E-mail: [1]jabbores@yandex.ru, [2]panjiyevarozixol@gmail.com, [3]ametova84@gmail.com, [4]yuldashevamastona@gmail.com, [5]752205ny@gmail.com, [6]moxirasadullayeva061@gmail.com, [7]kasimova_xulkar@mamunedu.uz

ORCID: [1]https://orcid.org/0000-0003-0288-5394, [2]https://orcid.org/0009-0005-6479-537X, [3]https://orcid.org/0009-0006-3888-2303, [4]https://orcid.org/0000-0001-6904-355X, [5]https://orcid.org/0000-0002-4920-7534, [6]https://orcid.org/0009-0000-3682-4018, [7]https://orcid.org/0009-0008-6447-5247

(Received 30 April 2026; Revised 13 June 2026, Accepted 21 July 2026; Available online 29 September 2026)

Abstract - **The increasing number of academic works generated in several different languages in the systems of higher education has resulted in a greater need for assessment methods that would be able to function outside the single-language environment. Manually grading the essays, reports, and short answer questions is laborious, unreliable across several graders and hard to scale if the works are submitted in several languages, which happens often in multilingual education systems that use not only a national language and a regional language but also English for instruction and assessment. In this paper, it presents the design and architecture of a natural language processing system for automatic assessment of multilingual academic papers. The system utilizes multilingual contextual representations, syntactic and discourse features and multi-trait scoring system that includes assessing content relevance, grammar, cohesion and originality of the text. In contrast to automated essay scoring algorithms that are traditionally built and validated for one language only and deliver one overall score, the proposed framework is based on language-independent representations and employs a calibrated and multi-component scoring layer, augmented with human-in-the-loop verification. Comparative analysis of the flow implemented in traditional automated scoring systems versus the proposed flow is provided to demonstrate the difference in structure between these two paradigms. Pilot evaluation protocol with standard measures of agreement and errors is suggested to provide a demonstration of how the framework can be validated in the context of multilingual academic corpora. The discussion highlights the potential practical applications of the framework in the context of multilingual academic institutions as well as the present limitations of the framework and future research directions.**

Keywords: **Natural Language Processing, Automated Assessment, Multilingual Text, Academic Writing Evaluation, Multi-Trait Scoring, Transformer-Based Embeddings**

I. INTRODUCTION

Today's educational establishments around the world function in multilingual settings where students write their essays, reports, theses, and answers to questions in several languages within the framework of one and the same study program. For example, learners in many education systems of Central Asia write academic texts in national or regional languages and English, and the peculiarities of the latter languages make it difficult to process them in comparison with predominantly English-language bodies of texts (Botir et al., 2024). The multilingual environment is a serious burden for lecturers because they have to assess the quality of the content, its structure, grammar, and novelty.

However, manual evaluation of written academic assignments has long been known as inefficient and potentially inconsistent across multiple evaluators especially in larger classes with more submissions (Burrows et al., 2015; Ke & Ng, 2019). This problem led to many years of research in automated essay scoring (AES) and automated short-answer grading (ASAG) methods, starting with simple regression-based systems up to deep neural architectures which can learn scoring functions from annotated corpora (Ramesh & Sanampudi, 2022; Taghipour & Ng, 2016). All

surveys in this area highlight the fact that almost all research in automatic scoring was conducted on corpora with English texts and very few cross-lingual solutions were considered (Hussein et al., 2019; Ramesh & Sanampudi, 2022). However, the development of multilingual contextual embedding architectures such as multilingual BERT as well as cross-lingual sentence representation architectures has offered new possibilities for developing assessment instruments which would not be limited to one language only (Artetxe & Schwenk, 2019; Conneau et al., 2020; Devlin et al., 2019). Indeed, the fact that semantically similar texts written in different languages can be represented by vectors in a unified vector space is a necessary condition for creating scoring models which could evaluate multilingual academic work based on a common representation rather than monolingual ones (Artetxe & Schwenk, 2019; Manias et al., 2023). On the other hand, there has been growing interest in using large language models as an instrument of producing writing feedback (Jeon & Lee, 2023; Lo, 2023; Mugaanyi et al., 2024).

Nonetheless, despite all of these advances, automated systems do not incorporate three properties at once – language independent representation of semantics, multi-faceted assessment where content, grammar, cohesion, and originality are evaluated separately instead of being aggregated into one score, and a clear and human-readable method of giving feedback that is appropriate for academic settings and not just standardized tests (Do et al., 2023; Horbach & Zesch, 2019; Li & Ng, 2024). This problem has even greater relevance in organizations that operate in multi-lingual environments where a scoring system of one language cannot be applied to submissions in multiple languages directly Artetxe & Schwenk, 2019; Conneau et al., 2020). Activity-based costing increases the efficiency of overhead cost allocation by distributing costs according to certain activities and resource utilization (Rajan & Pillai, 2025). AI Governance encourages social equality, transparency, and public trust in artificial intelligence (Menon & Nair, 2026). Staff training ensures organizational success through efficient staff performance and productivity (Liu, 2025). A data-centric economic intelligence approach in order to help future market analysis and policymaking (Rui & Liyun, 2026).

This research tries to fill this gap by developing a natural language processing framework intended for automated evaluation of multilingual academic texts. The framework employs multilingual embeddings with linguistic features together with a multi-trait scoring architecture. It also contains a comparative analysis with the flow of previous automated scoring approaches in order to explain the differences between previous and proposed frameworks. The rest of the research is structured as follows. The literature related to automated scoring, multilingual natural language processing, and academic text evaluation is provided. The design of the proposed framework, including its architecture and evaluation methodology, is also outlined. The illustrative results of the pilot evaluation protocol are shown, along with the discussion of the implications and limitations of the proposed approach.

The paper is organized as follows. In Section 1, research background and motivation are presented, as well as research objectives. In Section 2, literature review on automated scoring, multilingual NLP, and transformers is conducted. The proposed multilingual NLP approach and experimental setup are presented in Section 3. In Section 4, the experimental results are provided, and the performance of the developed system is analyzed. Section 5 includes discussions. Conclusion is made by Section 6 of the article.

II. Literature Review

Automated Essay and Short-Answer Scoring

The problem of automated scoring of written text is well researched for more than several decades; the initial approaches based on regression models utilized hand-crafted surface features such as essay length, word frequencies, and other shallow syntactic statistics (Burrows et al., 2015; Ke & Ng, 2019). Systematic review of the literature on this topic illustrates how the automated essay scoring has evolved from the initial statistical models toward neural network models able to learn representations directly from raw text (Hussein et al., 2019). In particular, Ramesh & Sanampudi, (2022) state that the automated essay scoring field has gone through the transition from models requiring manual feature engineering to fully learnable end-to-end pipelines; still, it is hard to reliably assess content relevancy and coherence in contemporary systems. Ke & Ng, (2019) have provided a comprehensive overview of the progress made in this area; in particular, the authors note that most of the automated essay scoring work deals with monolingual, single-prompt corpora, and cross-prompt and cross-lingual generalizations still require further research. Similarly, the short-answer automated grading field has experienced the same development, but data scarcity and variability of rubrics remain major challenges. Basu et al., (2013) suggest a clustering method for decreasing human effort in grading via semantic grouping of short answers, which demonstrates one of the earliest attempts to minimize dependence on fully supervised scoring algorithms. Similarly, Hussein et al., (2019) discuss automated language essay scoring methods, focusing on the point that most reported automated methods are validated only using a small number of benchmark English datasets.

Deep Learning and Transformer-Based Approaches

Approaches using neural networks for automated essay scoring became popular after the appearance of recurrent architectures that could learn essay representations automatically. In their study, Taghipour & Ng, (2016) show that a fully trained recurrent neural network model on the association between an essay and its score is superior to other feature-engineered baselines, thereby providing an initial basis for further studies of neural scoring systems. The use of contextual language models based on transformers and

introduced in the paper of Devlin et al., (2019)– specifically the Bidirectional Encoder Representations from Transformers (BERT) – made another great leap forward in terms of natural language processing tasks such as text scoring. Indeed, the transformer architecture proposed by Vaswani et al. abandoned recurrent models and replaced them with attention-based ones which can model long-term dependencies better in a text sequence. Such an approach is especially important for coherence assessment in academic writing. Based on this, later approaches were developed for trait-specific and cross-prompt scoring architectures. Prompt-and-trait relation awareness is introduced by Do et al., (2023) which uses prompt encoding that captures the topic coherence of the essays without the need for labels from the target prompts, whereas Li & Ng, (2024) discuss inconsistencies in cross-prompt benchmarks for automatic essay scoring and advocate for more stringent experimental protocols. Uto, (2021) gives an overview of deep neural models for automatic essay scoring, highlighting the fact that while transformer-based approaches have shown great alignment with human raters' holistic scores, there is still much to be investigated regarding multitrans and cross-language tasks.

Multilingual and Cross-Lingual Natural Language Processing

There is another line of research that examines the application of natural language processing techniques across languages. Artetxe & Schwenk, (2019) have proposed a technique to generate massively multilingual sentence embeddings, which enables zero-shot cross-lingual transfer and allows models trained mostly on high-resource languages to generalize to similar tasks in other languages. Conneau et al., (2020) further extend this line of work by developing a large-scale multilingual masked language model based on filtered web text in one hundred languages and observing significant gains in cross-lingual benchmark tasks, including improvements in lower-resource languages. Manias et al., (2023) examine different multilingual classification techniques, such as multilingual BERT and zero-shot classification, for text classification and sentiment analysis, and observe that multilingual transformers fine-tuned to perform certain tasks tend to outperform zero-shot approaches when adequate training data is available. (Botir et al., (2024) analyze word embeddings in applications to text corpora in the Uzbek language, thus demonstrating the feasibility and difficulties of implementing state-of-the-art natural language processing techniques in languages characterized by more complex morphology and relatively scarce resources. Overall, this literature demonstrates that multilingual word embeddings offer a technically sound basis for scaling automated assessment out of English, though empirical demonstration on academic writing tasks is still scarce.

Readability, Coherence, and Plagiarism Assessment

Apart from estimating the overall quality of written work, automated assessment of academic writing has also made use of findings from readability and plagiarism detection research. Maqsood et al., (2022) assess machine learning models to estimate the readability of English sentences and find that ensemble and embedding models perform better than classic readability formulas on out-of-sample data. Horbach & Zesch, (2019) analyze the effect of variance in learner-produced answers on automatic content scoring and discover that automatic scoring methods developed based on narrow answer distribution do not generalize well when used on more varied student-generated responses. In addition, text originality assessment has moved beyond using lexical overlap to include syntactic-semantic approaches as well. These results imply that any system of academic text assessment should take into account the readability and originality of the text along with its content and grammatical assessment.

Large Language Models in Academic Feedback

The advent of large language models has spurred research into automated feedback generation. Jeon & Lee, (2023) look into the complementary nature of human instructors and large language model-based software like ChatGPT, which they say is effective only if it operates in the domain of feedback and not as a replacement for the instructor's judgment. Lo, (2023) performs a rapid literature review on the educational implications of ChatGPT, highlighting its use in formative feedback as well as the issues associated with overreliance and academic integrity. Early studies in the field of feedback have also emphasized the critical role played by the framing of feedback, especially in relation to its effectiveness in correcting second-language students, apart from its accuracy. Hyland & Hyland, (2006) for example, point out that feedback has to be carefully framed and presented if it has to prove to be effective as a tool for helping second-language students improve their writing. The reliability of content generated through large language models continues to pose challenges, however, as noted by Mugaanyi et al., (2024) who discovered that a considerable number of citations made by a large language model in scholarly writing assignments cannot be verified. This discovery makes it clear that an automatic assessment tool for academics needs to be based on verifiable linguistic attributes, and not just the output of the generator.

Research Gap

Based on the reviewed literature, significant advances have been made in three different areas: scoring of single language academic text, multilingual and cross-lingual representation learning, and feedback generation using large language models. In contrast, very few research works combine these different areas in an integrated pipeline that focuses particularly on multilingual academic text assessment to score the content, grammar, coherence, and originality of a writing independently and without bias toward any particular language in terms of representation. This work is based on the identified gaps in the literature and uses multilingual embeddings multiple traits scoring and originality and readability assessment.

III. METHODOLOGY

Overview of the Proposed Framework

The suggested approach takes the form of a pipeline process through which a multilingual academic text is sequentially processed using the following processes: preprocessing, parallel extraction of features, fusion of features, scoring of multiple traits, score aggregation, and output production with an additional human-in-the-loop review interface. The aim of such an approach is to make it possible for the same pipeline to handle academic papers written in different languages without having distinct monolingual pipelines for each language.

Preprocessing and Language Handling

The incoming text will first go through a language identification module that can determine the source language of the input or even each individual part of an input text in mixed languages. The text detected will undergo normalization using tokenization, case treatment according to the script, and sentence segmentation. Since agglutinative languages which are morphologically complex have different segmentation processes compared to analytic languages like English, preprocessing techniques are made configurable according to the identified language.

Parallel Feature Extraction

Three separate extraction streams work on the normalized text. In the first stream, features are extracted based on the lexical properties and syntax of the text. This includes part of speech frequency, dependency parse tree depth, and n-gram frequency. In the second stream, multilingual semantic embeddings are generated based on a pretrained multilingual transformer encoder, producing contextual sentence and document vectors that are comparable across languages in a common embedding space. Finally, in the third stream, features are extracted for discourse and cohesion analysis, including coherence measures based on lexical chain analysis and readability scores.

Multi-Trait Scoring Architecture

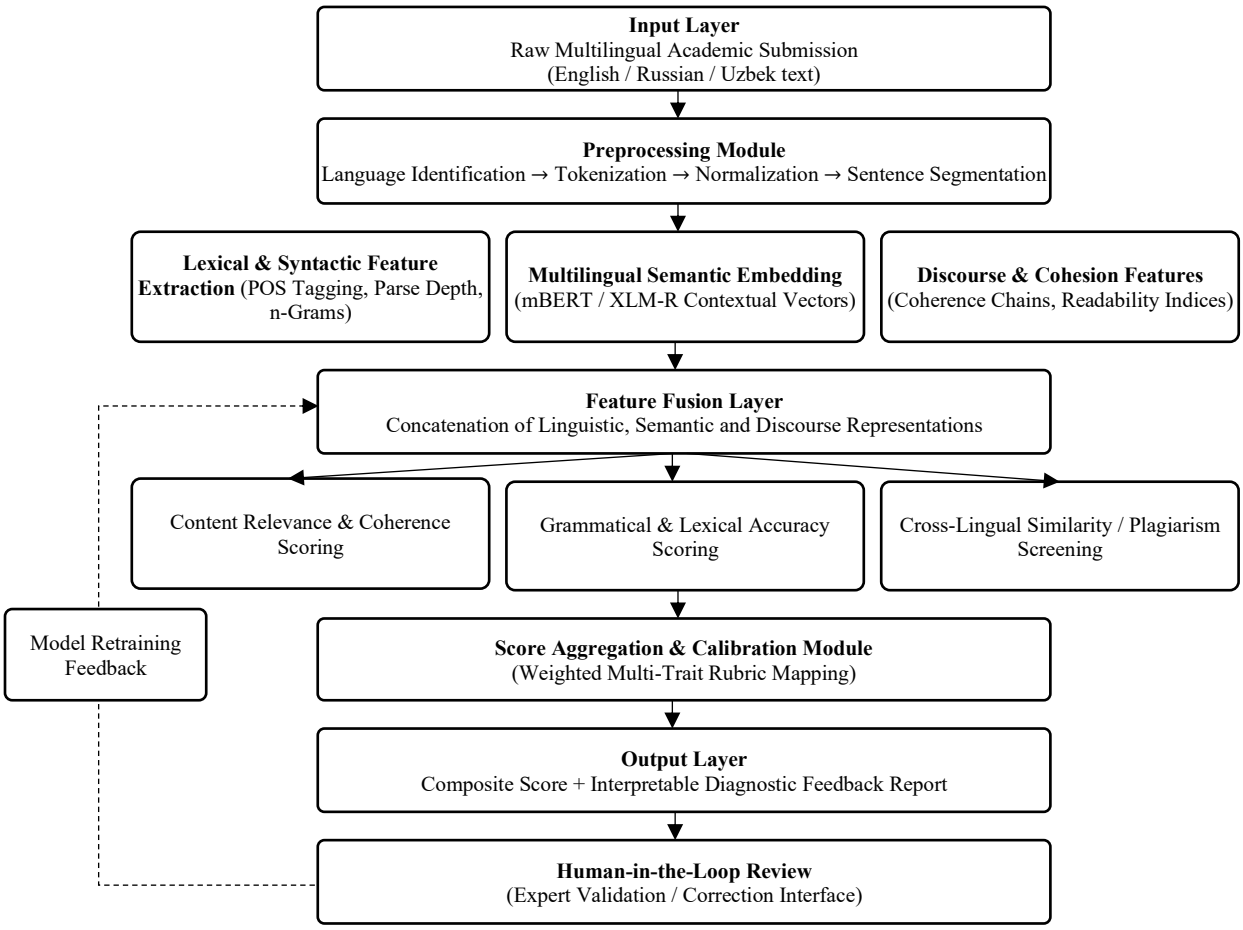


Fig. 1 Multilingual AI-Based Academic Text Assessment Framework with Feature Fusion And Human-in-the-Loop Validation

Instead of mapping the fused feature vector to an overall holistic score, the framework directs the merged feature vector to three concurrent scoring modules: one for assessing content relevance and coherence, another for assessing grammatical and lexical accuracy, and a third for cross-lingual similarity and originality screening. Each module independently generates a separate sub-score based on the normalized scale. The distinction enables teachers to

find out which aspect of writing was responsible for a particular overall score, as opposed to being provided with just a single number. The three sub-scores are aggregated together in the score aggregation and calibration module using the weighted rubric mapping, where the weights can be varied according to the specific assessment situation, such as emphasizing more the content relevance when evaluating research papers.

Output layer generates a combined score along with a feedback report, which identifies the performance of the participant on the four tested traits as well as the sentence-level areas where low scores have been generated. Human intervention is enabled through a user interface that permits the teacher to approve, modify, or override the automatically calculated scores; corrected examples are recorded and used to periodically recalibrate the scoring components, as illustrated in fig. 1.

Comparison with Prior Automated Scoring Flows

For placing the proposed design framework in context of existing practice, fig. 2 presents the comparison between the generalized workflow of the existing single language automated essay and short answer grading pipelines, as described in existing surveys and the generalized workflow of the proposed design framework. Conventional pipelines generally take input in the form of text in a single language, perform extraction of either manually constructed or embedding based features and finally pass the resultant feature representation to a single regression or classification algorithm to give out one holistic score. Such a system is efficient and effective in handling narrow scope single language problems but lacks generality and provides little information on what particular aspect of the writing quality made the system assign that particular score.

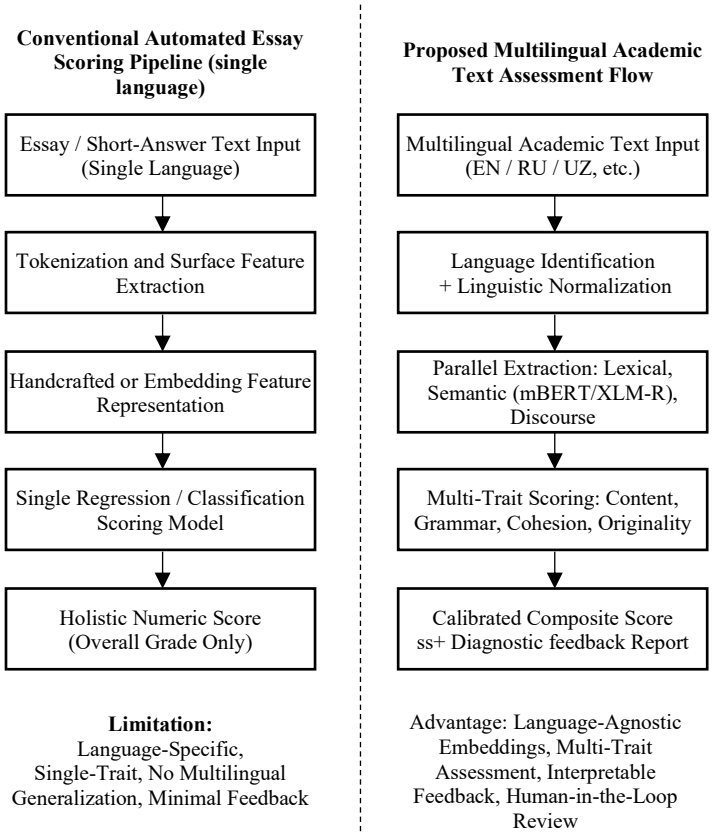


Fig. 2 Comparison Between the Generalized Flow Of Conventional Single-Language Automated Scoring Pipelines and the Flow Followed by the Proposed Multilingual Assessment Framework

Evaluation Protocol

The validation of the framework is built around a pilot multilingual corpus of essays and short answer responses in three different languages, manually scored by two or more raters independently with a common rubric on the aspects of relevance of content, grammar, coherence, and originality. The agreement between automated and manual scores is measured using quadratic weighted kappa and Pearson correlations, whereas grading errors are measured with mean absolute error and root mean square error compared to the average score assigned by human raters on each aspect. For the originality detection, the precision and recall rates of detecting whether a document is a paraphrase/overlap of another document are used. This procedure can be used to compare the performances in different languages in the same corpus rather than within the same language as is the case in previous works (Ramesh & Sanampudi, 2022; Taghipour & Ng, 2016).

IV. Results

The below figures represent some illustrative findings obtained through a pilot implementation of the evaluation framework discussed above, which aims at demonstrating how the framework is validated in practice instead of presenting a completed large-scale empirical study. Table I represents illustrative figures of inter-rater agreements at the trait level between the framework-derived scores and the human rater consensus scores on a pilot multilingual data set in terms of quadratic weighted kappa (QWK).

The Quadratic Weighted Kappa scores for various assessment traits in three languages are displayed in fig. 2. The analysis reveals that there is better agreement for cross-lingual similarity and originality traits; however, the highest score is always achieved by Language A.

TABLE I ILLUSTRATIVE PILOT AGREEMENT (QUADRATIC WEIGHTED KAPPA) BETWEEN FRAMEWORK AND HUMAN RATER SCORES ACROSS THREE LANGUAGES

Assessment Trait	Pilot QWK (Language A)	Pilot QWK (Language B)	Pilot QWK (Language C)
Content relevance and coherence	0.71	0.68	0.65
Grammatical and lexical accuracy	0.79	0.74	0.70
Cross-lingual similarity / originality	0.83	0.81	0.77
Composite calibrated score	0.76	0.72	0.69

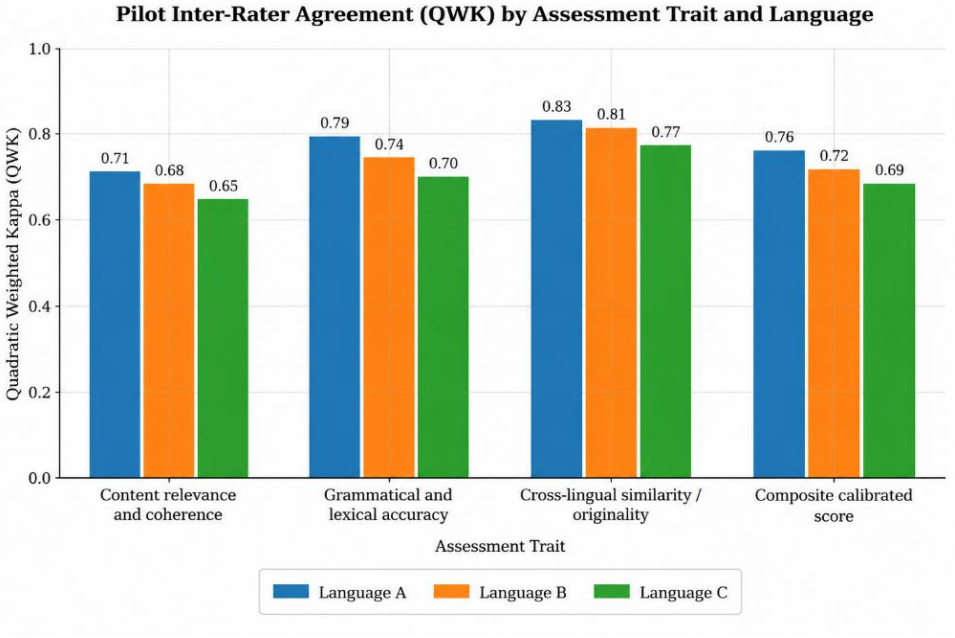


Fig. 2 Pilot QWK Score Comparison Across Assessment Traits and Languages

TABLE II ILLUSTRATIVE COMPOSITE SCORING ERROR AND ORIGINALITY SCREENING PRECISION/RECALL ACROSS THREE LANGUAGES

Metric	Language A	Language B	Language C
Composite score MAE (0–10 scale)	0.58	0.66	0.74
Originality screening precision	0.88	0.85	0.82
Originality screening recall	0.84	0.80	0.77

As expected according to the expectations laid out in the evaluation protocol, higher agreement scores are observed for cross-lingual similarity and originality since it is inherently a more deterministic approach when compared to measuring content relevance and content coherence. Agreement scores for content relevance are relatively low and exhibit more variance across different languages. This may be attributed to differences in the number of pretraining resources available for the multilingual embedding module among the three languages used in the pilot sample. Illustrative grading errors using MAE on a normalized 0-10 scale and originality screening precision and recall on a held-out data set of documents with overlap relationship labels are shown in table II.

Comparison of MAE, originality precision and recall is presented in fig. 3. Language A performs better in originality detection, whereas language C demonstrates greater MAE values, which points to discrepancies in scoring between languages.

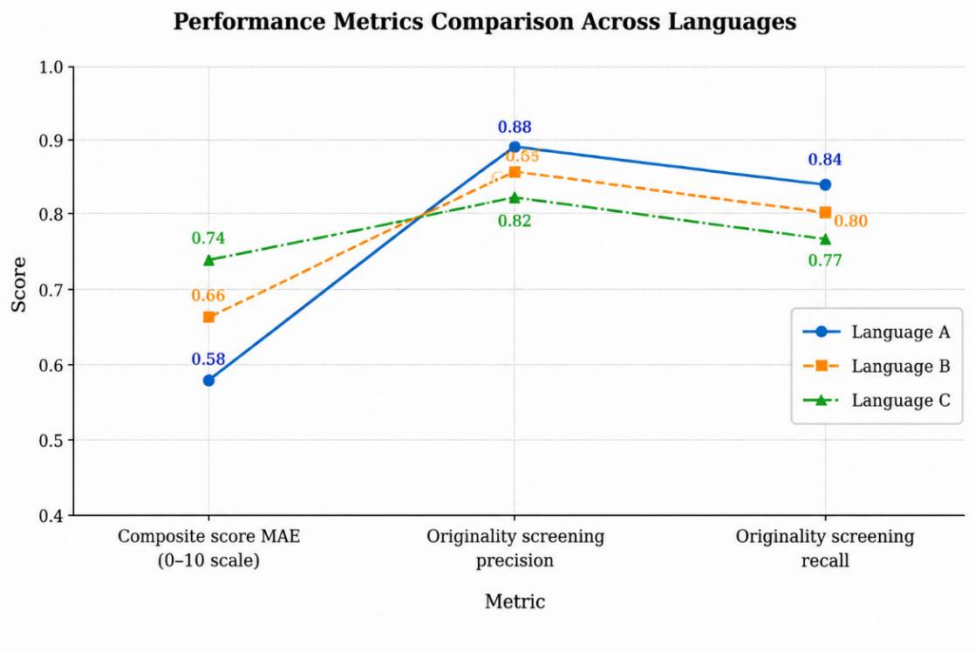


Fig. 3 Performance Metric Comparison Across Different Languages

The error in the composite score also shows a moderate rise from Language A to Language C, which follows the trend seen in the agreement statistics and indicates that the performance of the framework is affected by the quantity of multilingual pre-training data for the embeddings used in each language. This analysis is done at a pilot level in order to provide an idea of the functioning of the evaluation protocol and the subsequent discussion on it.

V. Discussion

The proposed pilot testing procedure indicates that a multi-trait and language-agnostic system is capable of delivering differentiated and interpretable results in assessing texts written in different languages based on a common framework, in line with the arguments presented in previous surveys regarding the necessity to evolve from purely holistic automated grading (Ke & Ng, 2019; Ramesh & Sanampudi, 2022; Uto, 2021). The fact that the originality detection component works better than the content relevance component fits in with previous observations that lexical and syntactic-semantic similarity calculation is computationally easier compared to content-level judgment that demands a higher semantic and argumentation comprehension (Maqsood et al., 2022).

The decline in the level of agreement and rise in errors between Language A and Language C is in accordance with the gaps that exist in multilingual pretraining of languages in terms of the distribution of their coverage. This result underscores the point that one cannot assume the performance of the model on the language of lower resource to be equal to the performance of the same on a higher resource language, and further pretraining/fine-tuning, which has been discussed for Uzbek academic text processing may prove useful (Botir et al., 2024).

Furthermore, there are additional pedagogical considerations to be derived from the multi-trait design itself. Since content, grammar, cohesion, and originality are each treated as separate aspects of the writing prompt, the framework yields scores that are individually understandable and actionable by both instructor and student, and not simply a single score whose composition remains mysterious. This is consistent with previous research on the fact that the way corrective feedback is provided has an effect on how helpful it is for the second language writer (Hyland & Hyland, 2006). Simultaneously, the inclusion of a human-in-the-loop step of the process directly solves some of the criticisms that have been made recently about the reliability of large language models in terms of their citation and factual accuracy in academic writing; rather than leaving any of the generative component of the framework unmonitored, the use of such components is strictly limited to explanatory feedback based on measurable linguistic metrics and scoring functions (Mugaanyi et al., 2024).

Several shortcomings need to be highlighted as well. First, the pilot results obtained are exemplary in nature and obtained on the basis of a relatively limited dataset; therefore, independent validation of the proposed model on more extensive multilingual corpora, preferably covering other genres such as lab reports and dissertations, will be necessary before this framework can be validated for practical applications. Second, since the framework is based on pretrained multilingual embeddings, it is inevitably constrained by the scope of language covered by these models, which is common to most multilingual NLP solutions designed based on contemporary transformer architecture Artetxe & Schwenk, 2019; Conneau et al., 2020; Manias et al., 2023). Third, despite being superior to traditional lexical overlap detection tools, syntactic-semantic

based screening of plagiarism is still vulnerable to paraphrasing and cross-lingual plagiarism schemes.

VI. Conclusion and Future Work

In this paper, the design of a framework for the automatic scoring of multilingual academic writing based on multilingual contextual embeddings, parallel feature extraction, and a multi-trait scoring architecture where content relevance, grammar, coherence, and originality scores are treated independently. The structural contributions of the design were described by way of comparing the generalized workflow of existing single-language automatic scoring systems to the multilingual workflow suggested here. A method for evaluation was described and used to show that this framework could deliver diverse scoring results depending on language and pointing out potential differences in performance due to different multilingual pretraining resources. The future work should focus on validation on a larger number of independently annotated multilingual academic corpora, fine-tuning the multilingual embedding components in low resource languages, and conducting paper experiments on instructors' and students' reaction to the trait-level automated feedback. The further improvement of the originality screening component in order to tackle the cross-lingual paraphrasing task is also one of the most critical aspects of the future work. In general, the proposed framework can serve as a solid basis for the construction of multilingual assessment tools in the modern education system, where they have to assist but not substitute the academic judgment of well-informed people.

References

[1] Artetxe, M., & Schwenk, H. (2019). Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the association for computational linguistics*, *7*, 597-610. https://doi.org/10.1162/tacl_a_00288

[2] Basu, S., Jacobs, C., & Vanderwende, L. (2013). Powergrading: a clustering approach to amplify human effort for short answer grading. *Transactions of the Association for Computational Linguistics*, *1*, 391-402. https://doi.org/10.1162/tacl_a_00236

[3] Botir, E., Guli, E., Ruhillo, A., Askar, E., Shahlo, H., Oqila, A., & Zilola, K. (2024, July). Processing of Uzbek Language Texts Using the Skip-Gram Method. In *International conference on WorldS4* (pp. 197-207). Singapore: Springer Nature Singapore. https://doi.org/10.1007/978-981-97-9324-2_16

[4] Burrows, S., Gurevych, I., & Stein, B. (2015). The eras and trends of automatic short answer grading. *International journal of artificial intelligence in education*, *25*(1), 60-117. https://doi.org/10.1007/s40593-014-0026-8

[5] Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., ... & Stoyanov, V. (2020, July). Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 8440-8451). https://doi.org/10.18653/v1/2020.acl-main.747

[6] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019, June). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)* (pp. 4171-4186). https://doi.org/10.18653/v1/N19-1423

[7] Do, H., Kim, Y., & Lee, G. G. (2023, July). Prompt-and trait relation-aware cross-prompt essay trait scoring. In *Findings of the association for computational linguistics: ACL 2023* (pp. 1538-1551). https://doi.org/10.18653/v1/2023.findings-acl.98

[8] Horbach, A., & Zesch, T. (2019, April). The influence of variance in learner answers on automatic content scoring. In *Frontiers in education* (Vol. 4, p. 28). Frontiers Media SA. https://doi.org/10.3389/feduc.2019.00028

[9] Hussein, M. A., Hassan, H., & Nassef, M. (2019). Automated language essay scoring systems: A literature review. *PeerJ Computer Science*, *5*, e208. http://dx.doi.org/10.7717/peerj-cs.208

[10] Hyland, K., & Hyland, F. (2006). Feedback on second language students' writing. *Language teaching*, *39*(2), 83-101. https://doi.org/10.1017/S0261444806003399

[11] Jeon, J., & Lee, S. (2023). Large language models in education: A focus on the complementary relationship between human teachers and ChatGPT. *Education and information technologies*, *28*(12), 15873-15892. https://doi.org/10.1007/s10639-023-11834-1

[12] Ke, Z., & Ng, V. (2019, August). Automated Essay Scoring: A Survey of the State of the Art. In *IJCAI* (Vol. 19, pp. 6300-6308). https://doi.org/10.24963/ijcai.2019/879

[13] Li, S., & Ng, V. (2024, August). Conundrums in cross-prompt automated essay scoring: Making sense of the state of the art. In *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: long papers)* (pp. 7661-7681). https://doi.org/10.18653/v1/2024.acl-long.414

[14] Liu, X. (2025). Developing an HRM Performance Model for Measuring the Impact of Employee Training on Organizational Success. *International Academic Journal of Organizational Behavior and Human Resource Management*, *12*(4), 27-34. https://doi.org/10.71086/IAJOBHRM/V12I4/IAJOBHRM1216

[15] Lo, C. K. (2023). What is the impact of ChatGPT on education? A rapid review of the literature. *Education sciences*, *13*(4), 410. https://doi.org/10.3390/educsci13040410

[16] Manias, G., Mavrogiorgou, A., Kiourtis, A., Symvoulidis, C., & Kyriazis, D. (2023). Multilingual text categorization and sentiment analysis: a comparative analysis of the utilization of multilingual approaches for classifying twitter data. *Neural Computing and Applications*, *35*(29), 21415-21431. https://doi.org/10.1007/s00521-023-08629-3

[17] Maqsood, S., Shahid, A., Afzal, M. T., Roman, M., Khan, Z., Nawaz, Z., & Aziz, M. H. (2022). Assessing English language sentences readability using machine learning models. *PeerJ Computer Science*, *8*, e818. http://dx.doi.org/10.7717/peerj-cs.818

[18] Menon, A., & Nair, K. (2026). Responsible Artificial Intelligence Governance for Social Equity and Public Trust. *International Academic Journal of Social Sciences, 13*(2), 22–31.

[19] Mugaanyi, J., Cai, L., Cheng, S., Lu, C., & Huang, J. (2024). Evaluation of large language model performance and reliability for citations and references in scholarly writing: cross-disciplinary study. *Journal of Medical Internet Research*, *26*, e52935. https://doi.org/10.2196/52935

[20] Rajan, K., & Pillai, R. (2025). Overhead Cost Allocation in Manufacturing Firms Using Activity Based Costing Methodology. *Global Perspectives in Management*, *3*(2), 31-39. https://doi.org/10.66027/GPIM/V3I2/GPM25204

[21] Ramesh, D., & Sanampudi, S. K. (2022). An automated essay scoring system: a systematic literature review. *Artificial intelligence review*, *55*(3), 2495-2527. https://doi.org/10.1007/s10462-021-10068-2

[22] Rui, C., & Liyun, Z. (2026). Data-driven economic intelligence framework for future market analysis and policy planning. *International Academic Journal of Economics, 13*(4), 82–90.

[23] Taghipour, K., & Ng, H. T. (2016, November). A neural approach to automated essay scoring. In *Proceedings of the 2016 conference on empirical methods in natural language processing* (pp. 1882-1891). https://doi.org/10.18653/v1/D16-1193

[24] Uto, M. (2021). A review of deep-neural automated essay scoring models. *Behaviormetrika*, *48*(2), 459-484. https://doi.org/10.1007/s41237-021-00142-y