

# RBASS: Rubric Based Automated Short Answer Scoring System

Ramesh Dadi<sup>1</sup><sup>a</sup> and Suresh Kumar Sanampudi<sup>2</sup>

<sup>1</sup>Department School of Computer Science and Artificial Intelligence, SR University, Warangal - 506371, Telangana, India

<sup>2</sup>Department of Information Technology JNTUH College of Engineering Jagtial, Nachupally, (Kondagattu), Jagtialdist Telangana, India

**Keywords:** Short Answer Scoring, Rubrics, Clustering, Content Scoring, Adversarial Responses.

**Abstract:** Automated short-answer scoring is a crucial tool in the educational system, enabling the quick and efficient assessment of students' responses. This process helps alleviate the challenges associated with manual evaluation while enhancing the reliability and consistency of assessments. However, it raises questions regarding whether it adequately considers the content and coherence of responses. Many researchers have tackled this issue and achieved promising results through natural language processing and deep learning advancements. Nevertheless, some researchers have focused solely on coherence evaluation, neglecting to test their models against adversarial responses, which limits the robustness of the model. This paper proposes a novel RBASS approach to evaluating answers based on coherence and content. Additional rubrics were incorporated into the existing dataset for content-based evaluation, resulting in optimal outcomes. The study comprehensively reviews and compares the system's performance using quantitative and qualitative metrics. Additionally, the model's performance is evaluated rigorously by training and testing it across multiple datasets and subjecting it to adversarial responses. The findings indicate that the model performs optimally both quantitatively and qualitatively, highlighting its effectiveness in assessing short-answer responses.


## 1 INTRODUCTION

Evaluating student performance constitutes a crucial aspect of the educational process. Nevertheless, the emergence of Massive Open Online Courses (MOOCs) and the expansion of class sizes have amplified the demand for an automated assessment framework. This system would ensure uniformity and precision in evaluating student responses, irrespective of quantity while delivering prompt feedback. Scholars have been dedicated to this domain since the 1960s, marked by the pioneering work of Page, E.B., who implemented the first automated assessment system (Zeng, Gasevic, et al. , 2023). Subsequently, researchers (Yao, Jiao, et al. , 2023) have delved into diverse feature extraction techniques and machine learning models.

Early researchers used statistical features such as TF-IDF, a bag of words, and N-grams (Kim, Lee, et al. , 2024). However, these statistical features and machine learning models were irrelevant to essay scoring. With the advancement of natural language

processing, feature extraction methods such as Word2vec (Kim, Lee, et al. , 2023) and GloVe (Chen, Li, et al. , 2020) were used in systems that could capture content by (Yang, Cao, et al. , 2020) but not semantics.

Researchers (Ridley, He, et al. , 2020), (Agrawal, and, Agrawal, 2018), (Cozma, Butnaru, et al. , 2018), (Jin, Wan, et al. , 2020) have used neural networks such as CNN, LSTM, and CNN+LSTM to capture sentence connections at the word-level embedding. However, these embedding methods could establish coherence only at the word level but not at the sentence level. To extract coherence from a prompt, researchers (Zhu, Sun, et al. , 2020), (Xia, Liu, et al. , 2019) have used transformer models such as USE, GPT2, and BERT, which extracted features sequentially and trained LSTM to capture sentence connectivity. In (Gaddipati, Nair, et al. , 2020) implemented a transformer based models extracted features with ELMO, and sum of word embeddings all the features and trained a sequential model. But with this approach coherence and content of the response will miss.

 <https://orcid.org/0000-0002-3967-8914>

However, while these models can embed prompts into vectors and train deep-learning models, they do not evaluate and test the responses in coherence, content, and cohesion parameters while assessing. Furthermore, these models require considerable human effort to label and require model training every time new prompts are added from different domains. Finally, these black box models are also prone to adversarial responses and cannot explain how they generate scores.

In conclusion, while automated assessment systems have the potential to offer consistency and accuracy, much work remains in developing systems that can accurately assess student responses in coherence, content, and cohesion parameters. From (Ding, Riordan, et al. , 2020), (Doewes and Pechenizkiy, 2021), (Horbach, Zesch, et al. , 2019), (Kumar, et al. , 2020), researchers must also address the issues of adversarial responses and explain ability to create more reliable and transparent assessment systems.

## 1.1 Organization

To outline the structure of the paper, we have organized the remaining sections as follows: Section 2 provides a discussion on the related work concerning text embeddings and deep learning models implemented in AES systems. Additionally, this section covers the challenges and limitations that arise in assigning a final score. In Section 3, we presented our dataset and the creation of rubrics irrespective of domain. Furthermore, in 3.3 and 3.4, we demonstrated our RBASS algorithm for the automated short-answer scoring system. In Section 4, we compare our experimental results with those of other models and demonstrate the performance of our model on adversarial responses. Finally, Section 5 discusses the conclusions drawn from our research and outlines future work.

## 1.2 Contribution

Here are the unique contributions of our paper.

- Integration of two distinct scoring metrics: coherence score and content score, providing a comprehensive evaluation of short answers.
- Development of the RBASS algorithm to provide content score, an unsupervised learning model in assessing responses across various domains.
- The model can be adapted to different subject areas, significantly reducing the reliance on human intervention.
- A rigorous comparative analysis between RBASS and other existing models will be presented, highlighting its consistency and robustness in short-answer evaluation.

## 2 RELATED WORK

Automated short-answer scoring is crucial in evaluating student responses to prompts, especially in assessing their understanding of domain-specific knowledge. Domain expertise and terminology are particularly significant because terms like "CELL" can have vastly different meanings in fields such as biology and physics. Consequently, evaluating student responses regarding domain-specific terminology poses a significant challenge within Natural Language Processing (NLP).

The journey of Automated Essay Scoring (AES) research traces back to its early stages in the 1960s and 1970s when rudimentary regression models and manually crafted feature extraction methods were utilized. Although these early attempts were essential, they laid the groundwork for subsequent advancements. As research progressed, AES incorporated statistical features like word count, word length, and word frequency (Li, Xi, et al. , 2023) and integrated machine learning models (Darwish, Darwish, et al. , 2020), (Rodriguez, Jafari, et al. , 2019).

The landscape of AES underwent a dramatic transformation with the emergence of natural language processing and neural networks. Feature extraction methods shifted towards automation, with word embedding techniques like word2vec gaining prominence for this purpose (Lun, Zhu, et al. , 2020), (Mathias, Bhattacharyya, et al. , 2018), (Song, Zhang, et al. , 2020). However, these methods had limitations, particularly in capturing semantic nuances and coherence in the evaluation process. Random, contextually unrelated words could easily mislead them.

Recent advancements have significantly shifted towards leveraging deep learning models instead of traditional machine learning approaches (Schlippe, Stierstorfer, et al. , 2022), (Dasgupta, Naskar, et al. , 2018), (Liu, Xu, et al. , 2019). While these deep learning models offer promising potential, they also present challenges. But recently researches (Künnecke, Filighera, et al. , 2024), (Süzen, Gorban, et al. , 2020), (Yang, Cao, et al. , 2020) some of the

researchers used transformer models and increased the performance of the model. One persistent issue revolves around the models' ability to distinguish between irrelevant responses and those demonstrating a solid grasp of word connectivity within the given domain. Moreover, these deep learning models' "black-box" nature raises several questions and concerns in the context of automated essay and short-answer scoring systems.

## 2.1 Confirming Relevance to the Prompt

Determining the relevance of a written response to a given prompt is a critical aspect of automated essay scoring. Recent research efforts (Chang, Kanerva, et al. , 2020) have explored various techniques to address this challenge. One approach involves embedding essays sentence by sentence using USE and sentence-BERT (Fernandez, Ghosh, et al. , 2022). And (Ridley, He, et al. , 2020) proposed a model to evaluate cross prompt essay, and used CNN, LSTM model, and extracted POS based features. While these methods aim to assess coherence and sentence-to-sentence connectivity, they may need to be revised to verify whether the response aligns with the prompt, is comprehensive, or incorporates appropriate domain-specific content. Other methods have also been employed, such as utilizing BERT (Mayfield, Black, et al. , 2020) for essay embedding and fine-tuning LSTM models. However, their primary focus has been on textual coherence rather than prompt relevance. (Yang, Cao, et al. , 2020) used two versions of BERT to extract features and train model but it is statistical approach.

## 2.2 Degree of Automation in Essay Scoring

The level of automation in automated essay scoring is a critical consideration. Most researchers (Taghipour, and, Ng, 2016), (Tay, Phan, et al. , 2018) have predominantly relied on supervised learning models to train student responses. However, it is worth noting that supervised learning models necessitate labeled data, which entails substantial manual human effort. Consequently, labeling data must be repeated when prompts or domains change (Yang, Cao, et al. , 2020). This dependency on human labeling diminishes the fully automated nature of the essay-scoring process.

Conveying Explanation and Assessing Completeness: Assessing the completeness of a response and conveying explanations for the given prompt is a multifaceted challenge. In some instances,

LSTM models (Kumar, Aggarwal, et al. , 2019) have been employed to evaluate sentence-to-sentence connectivity through context gates. These models summarize the content of sentence one and incorporate it into sentence two, with this process iteratively applied to subsequent sentences until reaching a final score. However, it is essential to note that such models primarily focus on the structural aspect of text and may not inherently address the critical question of prompt relevance. Furthermore, these models may struggle with adversarial responses, as their effectiveness is often contingent on the quality of text embeddings (Sawatzki, Schlippe, et al. , 2021).

In essence, the task of automated short answer scoring is a complex interplay between assessing structural coherence, prompt relevance, and the degree of human intervention required. Researchers continue to explore innovative methods to balance these aspects, striving to enhance the automation and effectiveness of this critical evaluation process.

# 3 METHODOLOGY

We have developed a new method for automated essay scoring using sentence-based text embedding to capture coherence from the response. We first collected key responses from at least four experts according to rubrics and added them to the dataset. Then, we tested our approach on two datasets - one standard and one domain-specific - consisting of 2300 responses from 600 students. We also evaluated the model's ability to handle different adversarial responses. Our proposed system utilizes the sentences BERT (Devlin, Chang, et al. , 2019) and RBASS model, as described in Figure 1 and Figure 2.

## 3.1 Dataset and Preprocessing

The ASAP Kaggle dataset, which includes 12,978 essays written by students in grades 8-10 in response to eight different prompts, was used to train our model. Two human raters evaluated each prompt, assessing 1500 or more essays for each prompt. Four prompts (Agrawal, and, Agrawal, 2018), (Kim, Lee, et al. , 2024), (Gaddipati, Nair, et al. , 2020), (Cozma, Butnaru, et al. , 2018) are source-dependent essays, while the others are not.

## 3.2 Features extraction using BERT

Sentence BERT is a sentence embedding technique that can convert text into vectors in a dynamic fashion, taking into account the context and

semantics of the text. Unlike other embedding techniques, such as word2vec and Glove, which convert text into vectors word-by-word, Sentence BERT can reconstruct the original sentence from the vector.

To use Sentence BERT, the text is tokenized into sentences, and each sentence is embedded into a 128-dimension vector using a pre-trained transformer model. For the ASAP datasets, the maximum number

Table 1 Sample essay vector after padding

dataset	Sample	Essay	dimension
<b>embedded vector by BERT</b>			
ASAP	[[	0.01323344 -	96*128
	0.09865774 0.032654971 ...		
	-0.0453132 -0.00221808		
	0.09867832]		
	[-0.02312332 -		
	0.07087123 -0.090920281		
	... -0.06158311 -0.09898121		
	0.07825336]		
	[-0.04417112 -		
	0.03577982 -0.012234987		
	... 0.01973515 0.09861238		
	0.00002164]		
	... [ 0. 0. 0.		
	0. ... 0. 0. 0. 0.		
	]]		

of sentences in an essay is 96 and 23, respectively. As a result, each essay is represented by 96\*128 vectors, respectively. Finally, all essays are padded to match the same dimensions of 96\*128 vectors. Table 1 portrays the sentence vectors for an essay.

## 4 RBASS ALGORITHM

The proposed approach aims to reduce the human effort required to evaluate student responses while providing scores based on content and completeness. This approach is unsupervised, meaning that it does not require labeled data. To begin, we collected responses from at least four experts for each score level, ranging from 0 to 5, for each prompt. Table 4 illustrates these rubrics. For a score of 5, the expert responses must be highly relevant to the prompt and consist of a sequence of sentences. For a score of 4, the expert responses are relevant and in sequence but may miss some points. For a score of 3, the expert responses are only partially relevant and may not be in sequence. For a score of 2, some points are relevant to the prompt, but not all, and the sequence may also be missed. For a score of 1, the expert response is a sentence that could be better explained. Finally, for a

score of 0, the expert responses are adversarial, meaning they do not provide any helpful information related to the prompt.

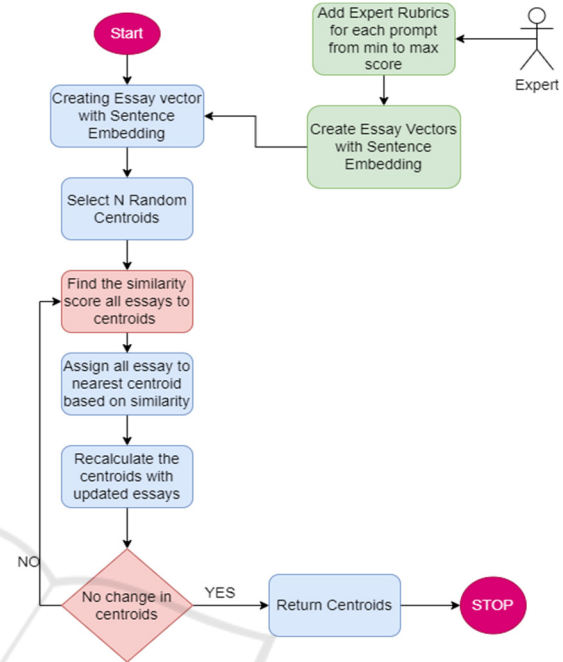


Figure 1 Working of the RBASS Algorithm

These expert responses are first embedded into vectors with Sentence\_BERT, then added to student responses. After embedding student and expert responses, all data is sent to the RBASS algorithm. The algorithm returns Centroids, as shown in table 6; the number of Centroids returned is the maximum mark allotted for the prompt. For example, if the maximum mark for a prompt is 5, then the algorithm returns six Centroids, including 0.

First, the algorithm selects default Centroids randomly, as shown in Fig 1. Then, it will find cosine similarity (1) between each response vector to each Centroids vector. Then, all responses will be assigned to the nearest Centroids list. After assigning all responses to the nearest Centroids, it will recalculate the Centroids, and this process will be repeated until there is no change in new and old Centroids.

$$\begin{aligned}
 \text{Similarity}(R_i, C_j) &= \cos(\theta) = \frac{R_i \cdot C_j}{\|R_i\| \cdot \|C_j\|} \\
 &= \frac{\sum_{i=0}^N R_i C_i}{\sqrt{\sum_{i=0}^N R_i^2} \sqrt{\sum_{i=0}^N C_i^2}} \quad (1)
 \end{aligned}$$

The score for a new student response is the Centroids id, to which the response is near. This score will be score-2 for the given response. However, the



LSTM (Poulton, and, Eliens, 2021) model will return the score-1.

This algorithm will assign score based on similarity that will consider content and completeness. If any type of adversarial response is found our algorithm can easily detects and assigns corresponding marks.

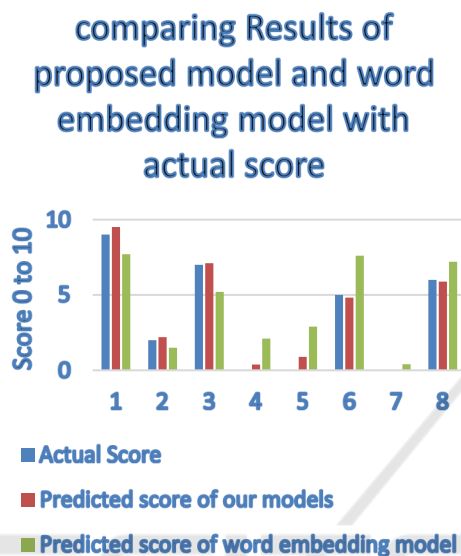


Figure 2 .comparison of Final score with word embedding, RBASS model with actual score of different test cases.

## 5 RESULT ANALYSIS

In Table 2, we provide a comprehensive overview of the content ratings achieved by the RBASS model, ranging from 0 to 5. Our evaluation process was meticulously designed, encompassing quantitative and qualitative analyses to ensure a thorough assessment. For the quantitative assessment, we calculated our model's average QWK score. Notably, our model outperformed all established models in this regard. The RBASS model achieved an impressive average QWK score of 0.796. This score attests to the model's exceptional capability in evaluating and rating content. We conducted a series of assessments on adversarial responses to evaluate the quality of our RBASS model's performance. These evaluations

included examinations of its proficiency in handling malicious responses while maintaining content and coherence.

Additionally, for a visual representation of our model's performance, we have included Figures 3 and 4. These figures provide a comparative scoring viewpoint between the RBASS model, word embedding models, and non-automated models. Figure 2 illustrates the prompt-wise comparison of all prescribed models to RBASS; in this, it is observed that our model performed consistently on all prompts. They visually represent the actual and predicted scores, effectively demonstrating our model's ability to consistently generate scores that closely align with the actual scores in every case. This remarkable accuracy sets the RBASS model apart from its counterparts, underscoring its robustness and trustworthiness in content evaluation. This RBASS model performed well when considering the content of the response.

## 6 CONCLUSIONS

In conclusion, our AES method combines coherence and content scoring metrics using an LSTM model for coherence and an RBASS model for content evaluation. By averaging these scores, we provide a comprehensive evaluation, demonstrating the effectiveness of our approach. Our experiments on the ASAP dataset showed that our model outperforms the state of the models, especially due to the RBASS model's domain-agnostic evaluation capability, reducing human effort. Additionally, we prepared adversarial test cases to test the model performance, in this our model RBASS and LSTM approach showed robustness when evaluating adversarial responses, such as those that intentionally mislead the model or those that are grammatically correct but semantically incorrect.

Furthermore, our model has proven its mettle, demonstrating excellent performance on both the ASAP and OS datasets. This success underscores its effectiveness in assessing essays and short answers based on content and coherence. Looking ahead, our future research will focus on trait-based AES systems that offer comprehensive quantitative and qualitative feedback for responses.

Table 2: Comparison of all prescribed models and proposed model Prompt wise QWK score on ASAP dataset.

Models	1	2	3	4	5	6	7	8	QWK
[8]	0.836	0.730	0.732	0.822	0.835	0.832	0.821	0.718	0.790
[27]	0.803	0.658	0.64	0.772	0.799	0.816	0.787	0.644	0.743
[37]	0.775	0.687	0.683	0.75	0.818	0.813	0.805	0.594	0.746
[36]	0.832	0.684	0.695	0.788	0.815	0.810	0.800	0.697	0.764
[41]	0.766	0.659	0.688	0.778	0.805	0.791	0.760	0.545	0.724
[2]	0.798	0.628	0.659	0.653	0.756	0.626	0.74	0.64	0.64
[7]	0.822	0.682	0.672	0.814	0.803	0.811	0.81	0.70	0.76
[43]	0.817	0.719	0.698	0.845	0.841	0.847	0.839	0.744	0.794
[2]	0.807	0.671	0.672	0.813	0.802	0.816	0.826	0.700	0.766
[39]	0.834	0.716	0.714	0.812	0.813	0.836	0.839	0.766	0.791
[17]	0.647	0.587	0.623	0.632	0.674	0.584	0.446	0.451	0.592
[24]	0.656	0.553	0.598	0.606	0.626	0.572	0.38	0.53	0.56
[30]	-	-	-	-	-	-	-	-	0.77
[21]	0.846	0.748	0.737	0.820	0.826	0.825	0.820	0.721	0.793
[1]	0.779	0.639	0.685	0.801	0.790	0.790	0.81	0.62	0.74
[4]	0.708	0.706	0.704	0.767	0.723	0.776	0.749	0.603	0.717
RBASS	0.823	0.715	0.711	0.819	0.822	0.821	0.811	0.691	0.796

In the future, we will continue our study on trait-based AES systems and test the model on more adversarial responses to test the robustness of the model. To handle Out of Vocabulary (OOV) words, we are creating a separate corpus related to the OS dataset domain to handle OOV words.

## REFERENCES

- Zeng, Z., Gasevic, D., & Chen, G. (2023, June). On the effectiveness of curriculum learning in educational text scoring. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 37, No. 12, pp. 14602-14610).
- Ridley, R., He, L., Dai, X., Huang, S., & Chen, J. (2020). Prompt agnostic essay scorer: a domain generalization approach to cross-prompt automated essay scoring. *arXiv preprint arXiv:2008.01441*.
- Agrawal, A., & Agrawal, S. (2018) Debunking Neural Essay Scoring.
- Do, H., Kim, Y., & Lee, G. G. (2024). Autoregressive Score Generation for Multi-trait Essay Scoring. *arXiv preprint arXiv:2403.08332*.
- Gaddipati, S. K., Nair, D., & Plöger, P. G. (2020). Comparative evaluation of pretrained transfer learning models on automatic short answer grading. *arXiv preprint arXiv:2009.01303*.
- Cozma, M., Butnaru, A. M., & Ionescu, R. T. (2018). Automated essay scoring with string kernels and word embeddings. *arXiv preprint arXiv:1804.07954*.
- Chang, L. H., Kanerva, J., & Ginter, F. (2022, July). Towards Automatic Short Answer Assessment for Finnish as a Paraphrase Retrieval Task. In *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)* (pp. 262-271).
- Cao, Y., Jin, H., Wan, X., & Yu, Z. (2020, July). Domain-adaptive neural automated essay scoring. In *Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval* (pp. 1011-1020).
- Schlippe, T., Stierstorfer, Q., Koppel, M. T., & Libbrecht, P. (2022, July). Explainability in automatic short answer grading. In *International conference on artificial intelligence in education technology* (pp. 69-87). Singapore: Springer Nature Singapore.
- [Dasgupta, T., Naskar, A., Dey, L., & Saha, R. (2018, July). Augmenting textual qualitative features in deep convolution recurrent neural network for automatic essay scoring. In *Proceedings of the 5th Workshop on Natural Language Processing Techniques for Educational Applications* (pp. 93-102).
- Ding, Y., Riordan, B., Horbach, A., Cahill, A., & Zesch, T. (2020, December). Don't take "nswvtnvakxpm" for an answer—The surprising vulnerability of automatic content scoring systems to adversarial input. In *Proceedings of the 28th International Conference on Computational Linguistics* (pp. 882-892).
- Darwish S.M., Mohamed S.K. (2020) Automated Essay Evaluation Based on Fusion of Fuzzy Ontology and Latent Semantic Analysis. In: Hassanien A., Azar A., Gaber T., Bhatnagar R., F. Tolba M. (eds) *The International Conference on Advanced Machine Learning Technologies and Applications*.
- Doewes, A., & Pechenizkiy, M. (2021). On the Limitations of Human-Computer Agreement in Automated Essay Scoring. *International Educational Data Mining Society*.

- Fernandez, N., Ghosh, A., Liu, N., Wang, Z., Choffin, B., Baraniuk, R., & Lan, A. (2022). Automated Scoring for Reading Comprehension via In-context BERT Tuning. arXiv preprint arXiv:2205.09864.
- Horbach A and Zesch T (2019) The Influence of Variance in Learner Answers on Automatic Content Scoring. *Front. Educ.* 4:28. doi: 10.3389/educ.2019.00028.
- Hussein, M. A., Hassan, H. A., & Nassef, M. (2020). A trait-based deep learning automated essay scoring system with adaptive feedback. *International Journal of Advanced Computer Science and Applications*, 11(5).
- Chen, Y., & Li, X. (2023, July). PMAES: Prompt-mapping contrastive learning for cross-prompt automated essay scoring. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 1489-1503).
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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*, pages 4171–4186.
- Kumar, Y., Aggarwal, S., Mahata, D., Shah, R. R., Kumaraguru, P., & Zimmermann, R. (2019, July). Get it scored using autosas—an automated system for scoring short answers. In *Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 33, No. 01, pp. 9662-9669)*.
- Kumar, Y. et al. (2020) “Calling Out Bluff: Attacking the Robustness of Automatic Scoring Systems with Simple Adversarial Testing.” ArXiv abs/2007.06796.
- Li, F., Xi, X., Cui, Z., Li, D., & Zeng, W. (2023). Automatic essay scoring method based on multi-scale features. *Applied Sciences*, 13(11), 6775.
- Liu, J., Xu, Y., & Zhu, Y. (2019). Automated essay scoring based on two-stage learning. arXiv preprint arXiv:1901.07744.
- Lun J, Zhu J, Tang Y, Yang M (2020) Multiple data augmentation strategies for improving performance on automatic short answer scoring. In: *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(09): 13389-13396
- Do, H., Kim, Y., & Lee, G. G. (2023). Prompt-and trait relation-aware cross-prompt essay trait scoring. arXiv preprint arXiv:2305.16826.
- Mathias S, Bhattacharyya P (2018) ASAP++: Enriching the ASAP automated essay grading dataset with essay attribute scores. In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*
- Mayfield, E., & Black, A. W. (2020, July). Should you fine-tune BERT for automated essay scoring?. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 151-162).
- Muangkammuen, P., & Fukumoto, F. (2020, December). Multi-task Learning for Automated Essay Scoring with Sentiment Analysis. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing: Student Research Workshop* (pp. 116-123).
- Ormerod, C. M., Malhotra, A., & Jafari, A. (2021). Automated essay scoring using efficient transformer-based language models. arXiv preprint arXiv:2102.13136.
- Künnecke, F., Filighera, A., Leong, C., & Steuer, T. (2024). Enhancing Multi-Domain Automatic Short Answer Grading through an Explainable Neuro-Symbolic Pipeline. arXiv preprint arXiv:2403.01811.
- Yao, L., & Jiao, H. (2023). Comparing performance of feature extraction methods and machine learning models in essay scoring. *Chinese/English Journal of Educational Measurement and Evaluation*, 4(3), 1.
- Rodriguez, P. U., Jafari, A., & Ormerod, C. M. (2019). Language models and automated essay scoring. arXiv preprint arXiv:1909.09482.
- Riordan, B., Flor, M., & Pugh, R. (2019, August). How to account for misspellings: Quantifying the benefit of character representations in neural content scoring models. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 116-126).
- Sawatzki, J., Schlippe, T., & Benner-Wickner, M. (2021, July). Deep learning techniques for automatic short answer grading: Predicting scores for English and German answers. In *International conference on artificial intelligence in education technology* (pp. 65-75). Singapore: Springer Nature Singapore.
- Poulton, A., & Eliens, S. (2021, September). Explaining transformer-based models for automatic short answer grading. In *Proceedings of the 5th International Conference on Digital Technology in Education* (pp. 110-116).
- Song, W., Zhang, K., Fu, R., Liu, L., Liu, T., & Cheng, M. (2020, November). Multi-stage pre-training for automated Chinese essay scoring. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 6723-6733).
- Süzen, N., Gorban, A. N., Levesley, J., & Mirkes, E. M. (2020). Automatic short answer grading and feedback using text mining methods. *Procedia Computer Science*, 169, 726–743
- 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).
- Tay, Y., Phan, M., Tuan, L. A., & Hui, S. C. (2018, April). SkipFlow: Incorporating neural coherence features for end-to-end automatic text scoring. In *Proceedings of the AAAI conference on artificial intelligence (Vol. 32, No. 1)*.
- Wang, Y., Wang, C., Li, R., & Lin, H. (2022). On the Use of BERT for Automated Essay Scoring: Joint Learning of Multi-Scale Essay Representation. arXiv preprint arXiv:2205.03835.
- Wang Z, Liu J, Dong R (2018a) Intelligent Auto-grading System. In: 2018 5th IEEE International Conference on

- Cloud Computing and Intelligence Systems (CCIS) p 430–435. IEEE.
- Zhu, W., & Sun, Y. (2020, October). Automated essay scoring system using multi-model machine learning. In CS & IT Conference Proceedings (Vol. 10, No. 12). CS & IT Conference Proceedings.
- Xia L, Liu J, Zhang Z (2019) Automatic Essay Scoring Model Based on Two-Layer Bi-directional LongShort Term Memory Network. In: Proceedings of the 2019 3rd International Conference on Computer Science and Artificial Intelligence p 133–137
- Yang, R., Cao, J., Wen, Z., Wu, Y., & He, X. (2020, November). Enhancing automated essay scoring performance via fine-tuning pre-trained language models with combination of regression and ranking. In Findings of the Association for Computational Linguistics: EMNLP 2020 (pp. 1560-1569).

