Received: 27 August 2023 | Revised: 23 November 2023 | Accepted: 20 December 2023

DOI: 10.4218/etrij.2023-0324

SPECIAL ISSUE

ETRI Journal WILEY

Dual-scale BERT using multi-trait representations for holistic and trait-specific essay grading

Minsoo Cho | **Jin-Xia Huang** | **Oh-Woog Kwon**

Language Intelligence Research Section, Electronics and Telecommunications Research Institute, Daejeon, Republic of Korea

Correspondence
Minsoo Cho, Language Intelligence Research Section, Electronics and Telecommunications Research Institute, Daejeon, Republic of Korea.
Email: mscho@etri.re.kr

Funding information
MSIT/IITP, Grand/Award Number: 2019-0-00004

Abstract

As automated essay scoring (AES) has progressed from handcrafted techniques to deep learning, holistic scoring capabilities have merged. However, specific trait assessment remains a challenge because of the limited depth of earlier methods in modeling dual assessments for holistic and multi-trait tasks. To overcome this challenge, we explore providing comprehensive feedback while modeling the interconnections between holistic and trait representations. We introduce the DualBERT-Trans-CNN model, which combines transformer-based representations with a novel dual-scale bidirectional encoder representations from transformers (BERT) encoding approach at the document-level. By explicitly leveraging multi-trait representations in a multi-task learning (MTL) framework, our DualBERT-Trans-CNN emphasizes the interrelation between holistic and trait-based score predictions, aiming for improved accuracy. For validation, we conducted extensive tests on the ASAP++ and TOEFL11 datasets. Against models of the same MTL setting, ours showed a 2.0% increase in its holistic score. Additionally, compared with single-task learning (STL) models, ours demonstrated a 3.6% enhancement in average multi-trait performance on the ASAP++ dataset.

KEYWORDS
automated essay scoring, deep learning methods, multi-task learning, multi-trait scoring, transformer-based models

1 | INTRODUCTION

Automated essay scoring (AES), which involves the use of computer systems to automatically evaluate essays in a manner similar to that of human graders, has been widely researched owing to its potential impact on educational assessment [1, 2]. Recognized for its practicality and cost effectiveness [3], AES research has evolved significantly from analyzing text complexity features [4] to leveraging machine-learning techniques [5], reflecting broader developments in technology and artificial intelligence.

Previous studies have employed various techniques, including traditional machine-learning approaches, such as linear regression and clustering. These techniques focus on using features extracted from essays, including average word length, paragraph count, and grammatical structure [2]. Certain studies [6] have combined feature engineering models with end-to-end models, demonstrating the improved potential of machine-learning AES approaches.

Deep neural networks, such as convolutional neural networks (CNNs) [7] and recurrent neural networks [8], have been employed to learn complex patterns effectively while eliminating the need for intricate feature engineering [9–11]. With the advent of bidirectional encoder representations from transformers (BERT) [12], these models have been significantly enhanced in terms of AES [13]. However, most relevant studies have concentrated on generating comprehensive holistic scores to evaluate essay quality. Although considerable human-like results have been observed in this area, there is a growing need to assess essays based on specific traits [14, 15], such as organization, content, and word choice, which differ across categories. Evaluating trait-specific attributes can provide detailed feedback and enrich the understanding and transparency of an essay's overall quality. Previous work [16] considered essay traits an auxiliary task, mainly concentrating on using their scores to improve holistic score prediction. However, this approach handles traits secondary to holistic accuracy. Consequently, the potential for delivering comprehensive feedback to students remains uninvestigated.

To overcome these limitations, we adopt a comprehensive approach that leverages the simultaneous prediction of holistic and multi-trait scores to establish a framework in which trait-specific representations directly contribute to accurate holistic score predictions. This interdependency enhances prediction accuracy for both aspects of essay evaluation, offering more precise and informative feedback. Importantly, our approach significantly boosts the explainability [17, 18] and interpretability of AES. By explicitly leveraging representations from multi-trait scores, our model not only provides improved predictions but also expresses the reasoning behind those predictions. This transparency empowers educators and students to understand how individual essay traits contribute to holistic assessment, thereby fostering a more effective learning process.

Furthermore, although transformer-based models, such as BERT, have shown significant potential for AES, particularly for holistic scoring [13], their capacity to perform comprehensive document-level modeling, especially for simultaneous multi-trait and holistic score predictions, remains relatively unexplored. Hence, we emphasize the critical need for an approach that not only integrates the power of transformers for holistic scoring, but also extends their capabilities to address the complexity of multi-trait evaluation.

To address these issues, our proposed DualBERT-Trans-CNN model introduces an innovative approach to AES. It adopts a hierarchical BERT structure as its first encoding scheme, which is strategically designed to leverage the strengths of transformer-based representations. This structure is adept at deriving sentence-level representations and subsequently constructing comprehensive document-level representations. Notably, this method captures the contextual relationships between sentences, making it particularly effective for long essays with complex content. This hierarchical BERT structure is complemented by a second encoding scheme that integrates a convolutional layer for local feature extraction on top of the BERT structure. This scheme extracts fine-grained local information from the essay, thereby enhancing the model's ability to assess and analyze trait-specific attributes. By incorporating both encoding approaches, our model combines the advantages of capturing the global context and local details, enabling a more comprehensive understanding of the essay content. Diverse experiments on the ASAP++ [19] and TOEFL11 [20] datasets demonstrate the efficacy of our model, which outperforms several baseline approaches, verifying its practicality and potential for advancing the field of AES.

Our main contributions can be summarized as follows:

- **Innovative unified model for holistic and multi-trait feedback**: We introduce DualBERT-Trans-CNN as a novel unified model that leverages a transformer-based architecture to provide simultaneous and accurate holistic and multi-trait feedback, with particular emphasis on trait-specific evaluations.
- **Hierarchical structure for real-world complexity**: The hierarchical structure of our model effectively encodes essays of varying lengths, thereby enabling it to address the diversity of real-world essays. Empirical tests on the ASAP++ dataset highlight its adaptability across diverse essay lengths and categories while establishing optimal sentence lengths.
- **Comprehensive validation and robustness analysis:** The efficacy of our model is confirmed through thorough validation, including baseline comparisons, extensive experimentation using the ASAP++ and TOEFL11 datasets, and detailed analyzes. These rigorous investigations reinforce the reliability of the proposed model, affirming its potential to provide trustworthy and insightful essay evaluations.

2 | RELATED WORK

2.1 | Holistic grading

Holistic AES grading has experienced substantial advancements in recent years, driven by the integration of deep-learning methods [21–23], which have proven effective in assessing both short and long essays, as

evidenced by various studies. In the domain of short essays [24, 25], deep learning techniques have been used to achieve holistic grading, whereas for longer essays, the methods presented in [10, 11, 13] yield significant results. Additionally, [15] utilized a simple neural network model enhanced with an attention mechanism and a hierarchical structure to augment its evaluation capabilities. Other studies [26, 27] have extended the use of neural networks by employing transformer-based pretrained language models for AES, demonstrating their effectiveness. The capabilities of GPT-3.5, particularly in AES and grading, were explored in a recent study [28].

Researchers have proposed innovative methodologies for generalizing essay scores across various prompts and domains [29–31]. For example, [32] presented a two-stage deep neural network that generated pseudo-data for both prompt-dependent and prompt-specific conditions, thereby enhancing the model adaptability across different prompts. The method proposed in [33] enhances the domain transferability of AES models, with the goal of developing models that can score essays effectively, regardless of the domain. However, many of these efforts have focused primarily on holistic scoring, with limited attempts to address trait-specific evaluation dimensions. In the context of these deep learning-based methodologies, our work extends this line of research by further advancing the capabilities of transformer-based architectures to encompass holistic and multi-trait evaluations.

2.2 | Utilizing specific traits

Some studies [34–36] have highlighted the importance of trait utilization, emphasizing the adoption of multi-task learning (MTL) approaches for multi-trait evaluation. In this context, [16] presented a thorough comparison between single-task learning (STL) and MTL, demonstrating the advantages of the latter. Numerous studies have investigated inherent trait characteristics to provide a deeper understanding and more accurate evaluation of essay quality. For example, [37] utilized discourse element identification at both the sentence and paragraph levels with an organizational evaluation to effectively assess argumentative essays. [38] proposed a hierarchical coherence model to evaluate the overall quality of documents by considering the coherence of an essay at varying granularities. Reference [39] employed an unsupervised learning method to obtain discourse-aware text representations, thereby enhancing the organization and argumentation strength of the AES. Furthermore, studies such as [40] have leveraged transformer models for text coherence assessment across multiple tasks.

2.3 | Generation of feedback comments

Research on essay evaluation is expanding beyond essay scoring to the generation of explanatory feedback [41–46], which is expected to aid writers in the direct development of their skills. Initial studies on these types of tasks predominantly addressed grammatical errors, with a common focus on the correct use of prepositions. A notable study [42] proposed a corpus of 1900 essays with detailed annotations of preposition errors and employed a neural retrieval–based method for feedback generation. However, this method is constrained by its inability to produce feedback beyond the training dataset. To address this rigidity, [43] proposed a more flexible hybrid model that combines neural retrieval-based techniques with a pointer generator network.

The feedback-comment-generation task gained further attention with the GenChal2022 shared task [44], which brought about the challenge of managing the diversity of freely generated comments. Hence, [45] suggested the use of generalized templates, which involves tagging grammatical errors using existing systems and replacing diverse comments with standardized templates. To expand the scope of feedback, [46] introduced a Chinese dataset featuring enriched commentary from narrative essays utilizing a two-stage planning method. These developments complement traditional scoring methods and offer an extended domain for AES.

3 | METHODS

3.1 | DualBERT-Trans-CNN model

As shown in Figure 1(A), DualBERT-Trans-CNN serves as a single-task encoder model that leverages a transformer-based hierarchical structure and, in Figure 1(B), represents an overall multitask learning framework that integrates $N + 1$ instances of the DualBERT-Trans-CNN model, where N denotes the total number of traits scored with an additional incorporated encoder to provide a holistic assessment of the essay. Each DualBERT-Trans-CNN model utilizes a document-level encoder consisting of two distinct encoding modules: BERT-TransEnc and BERT-CNN. Additionally, the prediction layer

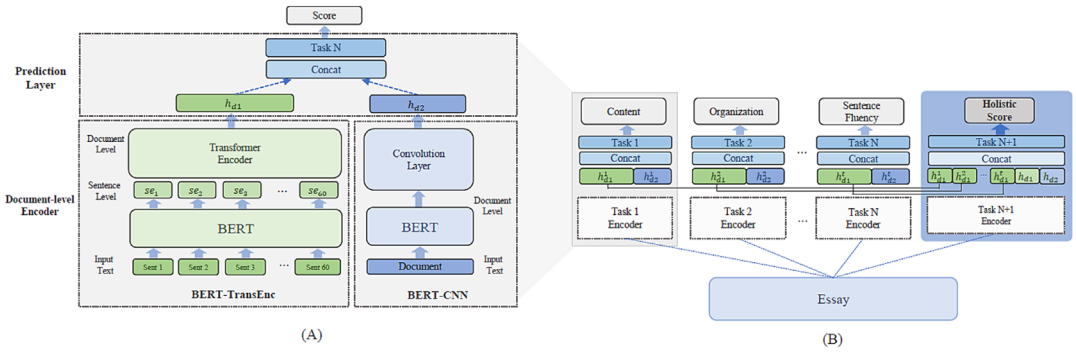


FIGURE 1 Architecture of the DualBERT-Trans-CNN model in STL and MTL settings (A) single-task encoder model (DualBERT-Trans-CNN) and (B) framework of multi-task learning.

merges the representations from the document-level encoder and predicts the final score of the essay. Detailed descriptions of the components are provided in the following subsections.

3.1.1 | Document-level encoder

BERT-TransEnc (sentence- and document-level)

The BERT-TransEnc module adopts a hierarchical structure for document-level encoding to capture and model the intersentence relationships that form the overall document structure by understanding how individual sentences interact with and relate to each other. Given an input document, D, consisting of sentences $X = (s_1, s_2, s_3, \ldots s_n)$, this module use BERT to generate a sentence-level representation, $se_i \in \mathbb{R}^d$, for each sentence.

$$SE = (se_1, se_2, \ldots, se_n). \tag{2}$$

By combining these sentence-level representations, we obtain a matrix, $SE \in \mathbb{R}^{n \times d}$, which represents the collective information of the sentences within the document. A transformer encoder layer is subsequently applied to the sentence-level representations to capture their dependencies and contextual information. We used a transformer encoder with $L = 5$ layers. The intermediate representations obtained in each layer are represented by the output, $h_i \in \mathbb{R}^{n \times d}$.

$$h = \text{TransformerEncoder}(SE), \tag{3}$$

$$H = (h_1, h_2, h_3, \ldots, h_L). \tag{4}$$

Finalthree layers of outputs are concatenated to form a new representation, $h_{\text{concat}} \in \mathbb{R}^{n \times 3d}$. This step captures essential contextual information from multiple layers and effectively encodes the overall semantics of the document.

$$h_{\text{concat}} = [h_{L-2}, h_{L-1}, h_L], \tag{5}$$

$$h_{d1} = \text{MeanPool}(h_{\text{concat}}). \tag{6}$$

Finally, a mean pooling operation processes the concatenated representation to generate a comprehensive document representation, $h_{d1} \in \mathbb{R}^{3d}$. This pooled representation encapsulates the key features from intermediate representations, yielding an efficient document-level representation that encodes the contextual relationships between sentences and captures the semantic essence of the document.

BERT-CNN (document-level)

The BERT-CNN provides another approach for document-level encoding. This method leverages a CNN to augment the final [CLS] output from the BERT, thereby optimizing the extraction of local features from a document. The final [CLS] output of document D, denoted as $X \in \mathbb{R}^d$, is subsequently processed by three distinct convolution filters with varying receptive field sizes (5, 10, and 15), each with a filter size of 256. The convolution filters are represented by the weight matrices, w_i, and bias terms, b_i. To incorporate nonlinearity, we apply a rectified linear unit (ReLU) function, f, to each convolution operation.

$$c_i = f(w_i \cdot X_{i:i+h-1} + b_i), \tag{7}$$

$$h_{d2} = \max(c_i). \tag{8}$$

The final representation, $h_{d2} \in \mathbb{R}^d$, is obtained by taking the maximum value over the feature maps, which serves as an integral component in subsequent prediction stages by complementing the first document-level representation provided by the BERT-TransEnc module.

Prediction layer. In an STL framework, where the objective is to predict individual traits or holistic scores, the

prediction layer of our model combines the final representations derived from the BERT-TransEnc and BERT-CNN modules, represented as h_{d1} and h_{d2}, respectively. These concatenated representations serve as inputs for generating predictive scores. For regression-based tasks, a sigmoid function is adopted to constrain the score range to zero and one, as shown in (9). Subsequently, the score is mapped to a specific range corresponding to each writing prompt, resulting in the final score.

$$y_{\text{pred}_{\text{trait}}} = \sigma\left(w_i \cdot \left[h_{d1}^j, h_{d2}^j\right] + b_i\right). \quad (9)$$

To predict the holistic score within the MTL framework, the trait-specific representations extracted from the BERT-TransEnc model, denoted as h_{d1}^j, where $j \in \{1, ..., T\}$ and T denotes the total number of traits) are concatenated with the document-level representations, h_{d1} and h_{d2}.The concatenated representation, $[h_{d1}^1, h_{d1}^2, ..., h_{d1}^T, h_{d1}, h_{d2}]$, is processed to derive a holistic score prediction, as shown in (10):

$$y_{\text{pred\_holistic}} = \sigma\left(w_i \cdot \left[h_{d1}^1, h_{d1}^2, ..., h_{d1}^T, h_{d1}, h_{d2}\right] + b_i\right). \quad (10)$$

This process enables the prediction of holistic scores by incorporating trait-specific representations from the BERT-TransEnc model and combining them with holistic document-level representations, thereby providing a richer understanding of the text. For multi-class classification tasks, where the level of an essay is classified as "low," "medium," or "high," the sigmoid functions in (9) and (10) are replaced by the Softmax function.

3.1.2 | Multi-task learning

For holistic score prediction within the MTL framework, we employ a weighted-loss approach to ensure a balanced contribution between trait-specific and holistic scores throughout the learning phase. We specifically utilize the mean squared error (MSE) loss, where the loss for both trait-specific and holistic tasks is represented by (11):

$$L = \text{MSE}\left(y_{\text{pred}}, y_{\text{true}}\right), \quad (11)$$

where y_{pred} denotes the output score of the model, and y_{true} denotes the corresponding ground-truth score. The final MTL loss is the weighted sum of individual losses. Greater weight is assigned to holistic score loss based on the findings of our empirical analysis.

$$L_{\text{MTL}} = \alpha * L_{\text{trait}} + (1-\alpha) * L_{\text{holistic}}. \quad (12)$$

In (12), α represents the weight assigned to the multi-trait loss within a prompt, whereas $(1-\alpha)$ represents the weight assigned to the holistic loss. In this study, we set α to 0.3 to more strongly emphasize holistic loss, which is consistent with our objective of capturing the comprehensive semantic context of a document. By employing this weighted approach, our model optimizes both trait-specific and holistic predictions.

4 | EXPERIMENTS

4.1 | Baselines

LSTM-CNN-att [47]: This is an integrated CNN with long short-term memory (LSTM) in a hierarchical sentence document framework. It utilizes attention pooling to capture crucial features of essay scoring.

SkipFlow LSTM [9]: This is a deep-learning architecture that embeds textual coherence modeling into AES by capturing semantic relationships within the hidden states of an LSTM.

Considering-Context-XLNet [48]: Based on XLNet, this model counters the bias in neural essay scoring caused by essay length and is designed to focus on content quality.

Trans-BERT-MS-ML-R [26]: Leveraging BERT for essay scoring, this model uses a joint learning method to generate multiscale essay representations and incorporates multiple loss functions alongside transfer learning techniques.

MTL-CNN-BiLSTM [16]: This model combines CNN-BiLSTM using an MTL approach with holistic scoring as the primary task and multi-trait scoring as an auxiliary task.

BERT [12]: The BERT model pretrains deep bidirectional representations from text and achieves state-of-the-art performance in numerous natural language processing tasks. We conducted experiments using the *bert-base* model with an added dense layer for MTL fine-tuning.

BigBird [49]: This transformer variant is designed to handle longer sequences using a sparse attention mechanism. In this study, we experimented with a *Bigbird-base* model.

4.2 | Datasets

Our model was trained and evaluated using two prominent essay-scoring datasets: ASAP++ [19] and TOEFL11 [20]. The ASAP++ dataset is an enhanced version of the widely recognized ASAP dataset [50] from the Kaggle competition in the AES domain and includes 12 978 essays distributed across three categories and eight prompts written by students in grades 7–10. Each essay was scored holistically based on multiple traits. Trait-specific scores cover detailed aspects of writing, such as "content," "organization," "word choice," and "sentence fluency." The TOEFL11 dataset originated from a collection of 12 100 vacation essays written by non-native English writers. Unlike ASAP++, TOEFL11 scores essays solely at the holistic level, classifying them into "low," "medium," and "high" categories. Table 1 provides detailed descriptions of the ASAP++ and TOEFL11 datasets. For a practical illustration, examples of essays from both datasets are provided in Figure A1 in the Appendix.

Additionally, to complement our cross-domain analysis for multi-trait prediction, we incorporated the ELLIPSE corpus [51] derived from the Kaggle competition and written by English language learners, which includes 3911 argumentative essays that offer trait-specific scores on "cohesion," "syntax," "vocabulary," "phraseology," "grammar," and "conventions."

4.3 | Evaluation metrics

We applied two prevalent evaluation metrics from the AES field: quadratic weighted kappa (QWK) [52] and accuracy. For the ASAP++ and ELLIPSE corpora, in which the task was to predict a score, we used QWK, which is suitable for regression problems. The QWK measures the agreement between two graders, represented by the predicted and actual essay scores, with values ranging from zero to one. For the TOEFL11 dataset, which corresponds to a classification task, we employed an accuracy metric to measure the proportion of correct predictions made by the model.

4.4 | Implementation details

For our model's BERT-TransEnc module, we leveraged the pretrained *bert-mini* model from Huggingface, which was designed with 256 hidden layers to process a maximum of 60 sentences in each essay, with each limited to 64 tokens. At the document level, a five-layer transformer with 786 hidden layers was employed. The BERT-CNN module utilizes a *bert-base* model with a size of 768, which is set to handle a maximum token length of 512 per essay.

Training was performed using fivefold cross-validation and repeated 10 times across both datasets. Specifically, for the ASAP++ dataset, we adhered to the fold ID from [10], with a data distribution of 60% for training, 20% for validation, and 20% for testing.

The key hyperparameters for model training included a learning rate of 2e−5, a dropout rate of 0.1, and a batch size of 16. Training was performed for a maximum of 30 epochs, with a patience value of 10, using an RTX 6000 NVIDIA GPU with 48 GB memory.

TABLE 1 Dataset descriptions of ASAP++ and TOEFL11.

Prompt	# Essays	Avg # word tokens	Category	Trait
Dataset: ASAP++				
1	1783	350	Persuasive	content/organization/word choice/sentence fluency/conventions
2	1800	350	Persuasive	content/organization/word choice/sentence fluency/conventions
3	1726	150	Source-based	content/narrativity/prompt adherence/language
4	1772	150	Source-based	content/narrativity/prompt adherence/language
5	1805	150	Source-based	content/narrativity/prompt adherence/language
6	1800	150	Source-based	content/narrativity/prompt adherence/language
7	1569	250	Narrative	content/organization/conventions/style
8	723	650	Narrative	content/organization/word choice/sentence fluency/conventions/voice
Dataset TOEL11				
1–8	12 100	348	Argumentative	-

4.5 | Results and analysis

4.5.1 | Performance comparison

Holistic score performance

Tables 2 and 3 present the holistic scoring performance of the proposed DualBERT-Trans-CNN model and existing state-of-the-art AES models on the ASAP++ and TOEFL11 datasets across two distinct settings. Table 2 presents the performance measures averaged over 10 five-fold cross-validation experiments, as suggested in [48]. Averaging across multiple runs ensures consistent and stable metric values by mitigating variations, thereby providing a reliable performance evaluation. As shown in Table 2, the Considering-Content-XLNet [48] model demonstrated higher performance than our model on the ASAP++ dataset. However, for the TOEFL11 dataset, our DualBERT-Trans-CNN model outperformed the considering-content-XLNet model by a substantial margin of 4.0%. This highlights the fact that although the Considering-Content-XLNet model is effective for analyzing the correlation between essay length and score, its performance was inconsistent across all datasets. In contrast, our model underscores its efficacy with both datasets, emphasizing its robustness in providing holistic score assessments.

A noteworthy observation from the performance comparisons in Table 2 is the apparent effectiveness of transformer-based models with neural networks. This is evident because both Considering-Content-XLNet and DualBERT-Trans-CNN, which are based on transformer architectures, outperformed models that rely on LSTM networks. This highlights the advantages of transformer-based architectures in capturing complex essay traits and patterns, leading to improved holistic scoring performance.

Table 3 complements Table 2 by presenting the performance of the essay scoring models across the STL and MTL settings. In this context, MTL refers to settings in which trait scores are utilized in a multitask learning manner. Table 3 presents the outcomes from a combination of 10 cross-validation experiments and one additional, thereby facilitating a holistic assessment of model capabilities. Notably, Trans-BERT-MS-ML-R [23] and MTL-CNN-BiLSTM [16] were performed solely within the scope of single-experiment score predictions.

In the context of STL, Trans-BERT-MS-ML-R [26] outperformed our DualBERT-Trans-CNN model by 0.7% on the ASAP++ dataset (Table 3). Although this achievement underscores the potential of the Trans-BERT-MS-ML-R model, the scope of this model is limited to a single holistic grading prediction, unlike our DualBERT-Trans-CNN model, which can handle both holistic and multi-trait outcomes. Hence, although our model yielded a lower performance from one perspective, it demonstrated versatility by extending its capabilities beyond the bounds of simple holistic scoring.

A comparison of the models in the MTL setting is particularly noteworthy, as listed in Table 3. To ensure fairness, we also present the performance of our model based on a single experiment that aligns with the approach used for other models. The results demonstrate the superiority of our DualBERT-Trans-CNN model over the MTL-CNN-BiLSTM model. Notably, the value of 2.0% underscores the efficacy of our model within the MTL framework.

Performance on long-sequence essays

Table 4[1] presents an evaluation of the holistic scoring performance for both long (averaging 450 tokens) and short (averaging 170 tokens) sequence essays across STL and MTL settings. In this evaluation, the DualBERT-Trans-CNN model was compared with the MTL-CNN-BiLSTM, Transformer-based BERT, and BigBird models. The token length for BERT was set to 512, whereas that for BigBird was set to 1024, considering GPU capacity constraints.

In this comparison, the DualBERT-Trans-CNN model outperformed the long-sequence essays in both STL and MTL settings, with a significant improvement of 2.1% in the MTL setting. These results highlight the robustness and efficiency of the proposed model in handling longer texts. These results also underscore the proficiency of our model, specifically the hierarchical BERT-TransEnc module, in its efficient processing and evaluation of long texts.

Multi-trait score performance

Table 5[2] presents detailed comparisons of the average multi-trait scores across all prompts for the ASAP++

TABLE 2 Performance comparison of holistic scoring results across ten cross-validation experiments.

Model	ASAP++ (QWK)	TOEFL (Acc)
LSTM-CNN-att (Dong et al., 2017)	0.764	0.667
SkipFlow LSTM (Tay et al., 2018)	0.764	-
Considering-Content-XLNet (Jeon et al., 2021)	0.786	0.728
DualBERT-Trans-CNN (Ours)	0.782	0.768

[1]The experiments for BERT and BigBird were implemented by the authors.

[2]In Table 5, the experiments for BERT were implemented by the authors. Specific numerical results for MTL-CNN-BiLSTM are absent due to the format of the data presentation in [16]. For a comparative visual comparison based on those bar graphs, refer to Figure 2.

TABLE 3 Performance comparisons across ten cross-validation experiments and a single cross-validation experiment highlighting STL and MTL settings.

MTL	Model	ASAP++ (QWK)	TOEFL (Acc)
×	LSTM-CNN-att (Dong et al., 2017)	0.764	0.667
×	SkipFlow LSTM (Tay et al., 2018)	0.764	-
×	Considering-Content-XLNet (Jeon et al., 2021)	0.786	0.728
×	Trans-BERT-MS-ML-R (Wang et al., 2022)	0.791	-
✓	MTL-CNN-BiLSTM (Kumar et al, 2022)	0.764	-
✓	DualBERT-Trans-CNN (Ours)	0.784	0.771

TABLE 4 Performance comparisons on long and short essays in the ASAP++ dataset.

MTL	Model	1,2,8 (long essays)	3,4,5,6,7 (short essays)
×	BERT (Devlin et al., 2018)	0.734	0.789
×	BigBird (Zaheer et al., 2020)	0.750	0.793
×	DualBERT-Trans-CNN (Ours)	0.753	0.790
✓	MTL-CNN-BiLSTM (Kumar et al, 2022)	0.740	0.778
✓	DualBERT-Trans-CNN (Ours)	0.761	0.794

TABLE 5 Performance comparisons on multi-trait feedback for the ASAP++ dataset.

Model	Cont	Org	WC	SF	Conv	PA	Lang	Narr	Style	Voice	Avg
LSTM-CNN-att (Dong et al., 2017)	0.707	0.649	0.621	0.612	0.605	0.731	0.640	0.699	0.659	0.544	0.647
BERT (Devlin et al., 2018)	0.719	0.668	0.676	0.625	0.656	0.731	0.659	0.703	0.693	0.610	0.674
DualBERT-Trans-CNN (Ours)	0.726	0.682	0.680	0.663	0.668	0.732	0.660	0.704	0.693	0.621	0.683
Δ (%)	1.9	3.3	5.9	5.1	6.3	0.1	2.0	0.5	3.4	**7.7**	3.6

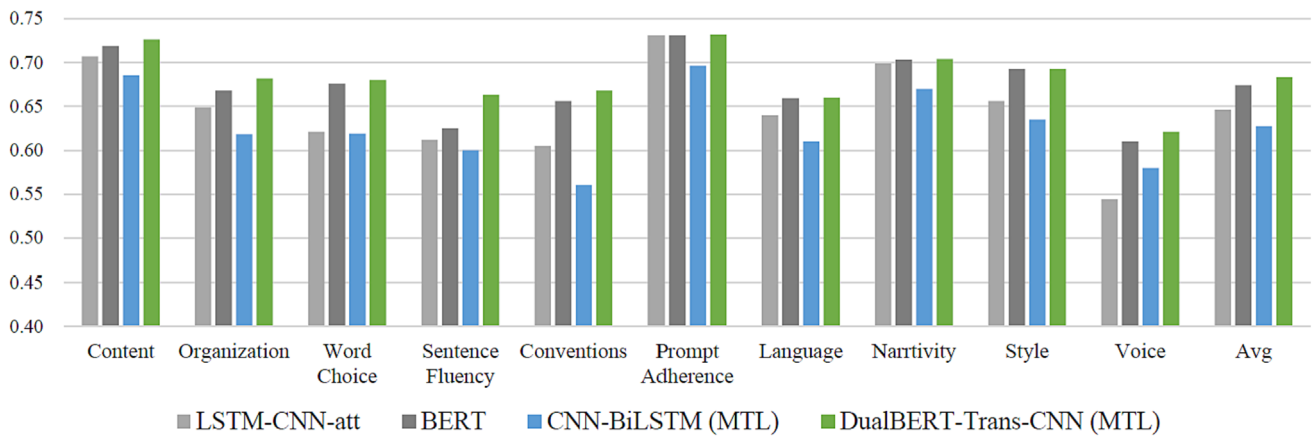


FIGURE 2 Performance comparisons on multi-trait scoring of ASAP++ dataset including the CNN-BiLSTM (Kumar et al., 2022) results.

dataset. The associated visual representations are shown in Figure 2. The LSTM-CNN-att model, which was setup as an STL system, was specifically trained to independently predict individual traits. In contrast, the other models predicted both trait and holistic scores simultaneously.

As shown in Table 5, our proposed DualBERT-Trans-CNN model provided robust performance for multi-trait scoring, with an average improvement of 3.6% across all traits when compared with the lowest scores of each. This is particularly evident for the traits of "voice," "conventions," "word choice," and "sentence fluency,"

where it significantly outperformed the LSTM-CNN-att model. These improvements highlight the advantages of incorporating pretrained information from transformer-based models into both BERT [12] and our proposed model, effectively capturing the essential elements of writing traits. Furthermore, our analysis revealed that the traits with the most pronounced enhancements largely aligned with those from the "persuasive" and "narrative" categories, as described in Table 1. Furthermore, a comparative assessment between the LSTM-CNN-att and the other models indicated the superiority of the MTL framework over STL, which can be attributed to integrated trait information processing during the training phase. These results demonstrate the applicability of the DualBERT-Trans-CNN model for grading holistic and trait-specific scores in the MTL framework within these essay categories.

4.5.2 | Ablation test results

The ablation study summarized in Table 6 provides a comprehensive understanding of the effects of each component within the DualBERT-Trans-CNN model on its overall efficacy. As described in Section 3.1.1, BERT-TransEnc and BERT-CNN are document encoders that utilize a single BERT model, whereas DualBERT-Trans utilizes a dual-scale BERT approach without a convolution layer and is a simplified version of the proposed DualBERT-Trans-CNN model.

A comparative analysis revealed the advantages of dual-scale BERT models over their counterparts that rely on a singular BERT model, such as the BERT-TransEnc and BERT-CNN. The DualBERT-Trans model, which utilizes a dual-scale BERT without a convolution, showed a 2.2 and 1.8% higher holistic scoring performance than BERT-TransEnc and BERT-CNN, respectively. This intermediately demonstrates an improved understanding of essays by capturing both sentence- and document-level information, which is essential for recognizing coherence and logical structure. The subsequent inclusion of a convolutional layer in DualBERT-Trans-CNN further refined this approach by extracting local features, thereby enhancing the model's ability to interpret and score essays of varying lengths and complexities. Moreover, when considering multiple traits within an MTL framework, DualBERT-Trans-CNN demonstrated superior performance on STL, underscoring the significance of utilizing multi-trait data within an MTL framework for comprehensive essay assessment. Overall, integrating document encoding techniques, embedding a convolutional layer in the BERT-CNN, and deploying a multi-trait-guided MTL framework significantly improved the performance of our model for both holistic and multi-trait essay scoring.

4.5.3 | Further analysis

Computational resource analysis

Table 7 presents a comparative analysis of the parameter sizes and runtime efficiencies of the DualBERT-Trans-CNN and BERT-based modules under holistic and multi-trait learning conditions. We assessed the total

TABLE 6 Performance comparisons of DualBERT-Trans-CNN modules on the ASAP++ dataset.

MTL	Model	1	2	3	4	5	6	7	8	Avg (holistic)	Avg (multi-trait)
✓	BERT-TransEnc	0.826	0.696	0.658	0.779	0.804	0.792	0.812	0.710	0.760	0.687
✓	BERT-CNN	0.799	0.660	0.688	0.803	0.805	0.812	0.823	0.723	0.764	0.689
✓	DualBERT-Trans	0.824	0.703	0.686	0.806	0.811	0.817	0.830	0.748	0.778	0.694
✓	DualBERT-Trans-CNN	0.822	0.710	0.689	0.808	0.813	0.827	0.832	0.752	0.782	0.697
×	DualBERT-Trans-CNN	0.824	0.694	0.684	0.806	0.810	0.820	0.830	0.742	0.776	-

TABLE 7 Parameter and runtime comparison for DualBERT-Trans-CNN and BERT-based modules during scoring on prompt 1.

MTL	Model	Parameters (M)	Training time (s)	Inference time (s)
✓	BERT-TransEnc	168	633	1.73
✓	BERT-CNN	224	1091	3.44
✓	DualBERT-TransEnc	277	1570	4.44
✓	DualBERT-Trans-CNN	313	1702	4.84

training time over 30 epochs and the inferencing time on the test set using the ASAP++ dataset for Prompt 1 without early stopping. Although the DualBERT-Trans-CNN model suffered more parameters, extended training, and longer inference times, these increases were quite modest considering the performance gains listed in Table 6. The moderate rise in resource usage by the DualBERT-Trans-CNN model was proportional to its advanced capabilities, underscoring the efficiency in processing and managing the complexities of the dataset, particularly within an MTL framework.

Effects of the loss weight ratio

Table 8 presents the effects of adjusting the weight loss ratio between holistic and multi-trait losses on the performance of the DualBERT-Trans-CNN model. The table shows that the performance changed with the loss weight distribution. A loss weight ratio of 0.3 for multi-trait loss and 0.7 for holistic loss yielded the best average performance for holistic scoring. This ratio represents the model's preference for a greater emphasis on holistic loss during learning. These empirical findings indicate that our model is more influenced by the holistic loss component than by the multi-trait loss component during training, suggesting that our model emphasizes the holistic aspect when considering these two elements during the essay-grading task. Furthermore, because the model was designed to leverage the representation of multi-trait information by concatenating its representations when predicting holistic loss, it inherently emphasized the training objective of multi-trait scoring, further underlining the importance of holistic scoring.

Effects of sentence length and count

Tables 9 and 10 present the effects of sentence tokenization on the proposed model. Specifically, they revealed how variations in sentence token length and the total number of encoded sentences in an essay influenced model performance. Table 9 details how variations in token length affected the performance of both our model and the hierarchical LSTM-CNN-att model [47]. By applying the same hyperparameter settings to the LSTM-CNN-att model as used with our DualBERT-Trans-CNN, we ensured a precise comparison. The results revealed that differences in token length minimally affected the performance of both models, with a slight improvement observed for longer-sentence tokens for the LSTM-CNN-att model.

Table 10 lists the number of sentences from the essay that affect model performance. The results show that essays with 60 encoded sentences outperformed those with 30 sentences by 2% in terms of holistic scoring. This suggests that the number of sentences is more important than the sentence length. Ensuring adequate representation of an essay, particularly in the context of the number of encoded sentences, is crucial for a comprehensive and accurate evaluation.

Performance in the cross-domain setting

Table 11 presents the performance of the DualBERT-Trans-CNN model compared with the LSTM-CNN-att model in cross-domain settings, specifically utilizing the ELLIPSE corpus and ASAP++ dataset. When trained on ELLIPSE and tested on ASAP++'s Prompt 1 argumentative essays, the DualBERT-Trans-CNN model outperformed the LSTM-CNN-at model in both multi-trait and holistic scoring. Conversely, when trained on ASAP++

TABLE 8 Effects of ratio of loss weight on holistic and trait loss.

Loss weight (trait, holistic)	Holistic (QWK)	Multi-trait (QWK)
0.3, 0.7	0.782	0.697
0.5, 0.5	0.780	0.697
0.7, 0.3	0.777	0.697

TABLE 10 Impact of varying numbers of encoded sentences in an essay on the proposed model on the ASAP++ dataset.

Number of sentences	Holistic (QWK)	Multi-trait (QWK)
30	0.762	0.689
60	0.782	0.697

TABLE 9 Impact of varying sentence token lengths on the ASAP++ dataset.

	Sentence token length	Holistic (QWK)	Multi-trait (QWK)
DualBERT-Trans-CNN (Ours)	32	0.780	0.696
	64	0.782	0.697
	128	0.781	0.696
LSTM-CNN-att (Dong et al., 2017)	32	0.738	0.636
	64	0.745	0.644
	128	0.745	0.645

TABLE 11 Performance comparisons on cross-domain setting for ASAP++ and ELLIPSE corpus dataset.

Train → test	Model	Cont	Org	WC	SF	Conv	Multi-trait (Avg QWK)	Holistic (QWK)
ELLIPSE corpus → ASAP++	LSTM-CNN-att (Dong et al., 2017)	0.386	0.436	0.436	0.391	0.423	0.420	0.386
	DualBERT-Trans-CNN (Ours)	0.417	0.471	0.391	0.423	0.420	0.474	0.421
Train → test	**Model**	**Cohesion**	**Vocab**	**Syntax**	**Conv**		**Multi-trait (Avg QWK)**	**Holistic (QWK)**
ASAP ++ → ELLIPSE corpus	LSTM-CNN-att (Dong et al., 2017)	0.334	0.329	0.284	0.329		0.319	0.334
	DualBERT-Trans-CNN (Ours)	0.400	0.318	0.414	0.421		0.388	0.305

and tested on ELLIPSE, the DualBERT-Trans-CNN model scored higher on multi-trait scoring but not on holistic scoring. This indicates that, although our model demonstrates the potential for domain generalization in trait-specific scoring, there is room for improvement in holistic scoring in diverse settings, including few- or zero-shot learning cases.

5 | CONCLUSIONS

In this study, we introduced the novel DualBERT-Trans-CNN model: a transformer-based framework powered by dual-scale BERT encoding that ensures diverse document-level representations and exhibits adaptability across various essay lengths. Our rigorous evaluation process, which encompassed baseline comparisons, diverse experimental setups, and ablation tests, demonstrated the superior robustness and validity of our approach. The analysis performed on the ASAP++ and TOEFL11 datasets strongly demonstrated the proficiency of our model in trait-specific scoring, emphasizing its significance in the AES domain.

In the future, we aim to advance the AES field by integrating more actionable and interpretable feedback mechanisms, considering the emergence of technologies, such as GPT-4 and similar models. Despite challenges in scoring consistency and controllability, large language models (LLMs) offer promising opportunities for augmenting AES through enriched feedback and dataset curation. Our goal is to synergize our model's precision with the generative capabilities of LLMs to improve assessment accuracy and feedback quality, thereby contributing to educational enhancement.

ACKNOWLEDGMENTS

This work was supported by Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (2019-0-00004, Development of semi-supervised learning language intelligence technology and Korean tutoring service for foreigners).

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

ORCID

Minsoo Cho https://orcid.org/0009-0002-3847-2022
Jin-Xia Huang https://orcid.org/0000-0002-2851-5124

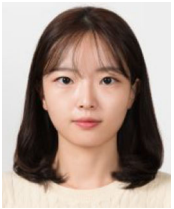
REFERENCES

1. P. W. Foltz, D. Laham, and T. K. Landauer, *The intelligent essay assessor: applications to educational technology*, Interact Multimed. Elecron. J. Comput-Enhanced Learn. **1** (1999), no. 2, 939–944.
2. Y. Attali and J. Burstein, *Automated essay scoring with e-rater® V. 2*, J. Tech. Learn. Assessment **4** (2006), no. 3.
3. E. B. Page, *The imminence of... grading essays by computer*, Phi Delta Kappan **47** (1966), no. 5, 238–243.
4. L. M. Rudner and T. Liang, *Automated essay scoring using Bayes' theorem*, J. Tech. Learn. Assessment **1** (2002), no. 2.
5. V. V. Ramalingam, A. Pandian, P. Chetry, and H. Nigam, *Automated essay grading using machine learning algorithm*, J. Phys. Conf. Ser. **1000** (2018), DOI 10.1088/1742-6596/1000/1/012030
6. J. Liu, Y. Xu, and Y. Zhu, *Automated essay scoring based on two-stage learning*, arXiv preprint, 2019, DOI 10.48550/arXiv.1901.07744
7. K. O'Shea and R. Nash, *An introduction to convolutional neural networks*, arXiv preprint, 2015, DOI 10.48550/arXiv.1511.08458
8. A. Sherstinsky, *Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network*, Phys. D: Nonlin. Phenom. **404** (2020), DOI 10.1016/j.physd.2019.132306
9. Y. Tay, M. Phan, L. A. Tuan, and S. C. Hui, *SKIPFLOW: incorporating neural coherence features for end-to-end automatic text scoring*, Proc. AAAI Conf. Artif. Intell. **32** (2018), no. 1.
10. K. Taghipour and H. T. Ng, *A neural approach to automated essay scoring*, (Proceedings of the 2016 conference on empirical

methods in natural language processing, Austin, TX, USA), 2016, pp. 1882–1891.
11. D. Alikaniotis, H. Yannakoudakis, and M. Rei, *Automatic text scoring using neural networks*, arXiv preprint, 2016, DOI 10.48550/arXiv.1606.04289
12. J. Devlin, M. W. Chang, K. Lee and K. Toutanova, *BERT: Pre-training of deep bidirectional Transformers for language understanding*, arXiv preprint, 2018, DOI 10.48550/arXiv.1810.04805
13. V. J. Schmalz and A. Brutti, Automatic assessment of English CEFR levels using BERT embeddings, (Proceedings of the Eighth Italian Conference on Computational Linguistics, Accademia University Press), 2021, pp. 293–299.
14. M. A. Hussein, H. A. Hassan, and M. Nassef, *A trait-based deep learning automated essay scoring system with adaptive feedback*, Int. J. Adv. Comput. Sci. Appl. **11** (2020), no. 5, DOI 10.14569/IJACSA.2020.0110538
15. S. Mathias and P. Bhattacharyya, *Can neural networks automatically score essay traits?* (Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications, Seattle, WA, USA), 2020, pp. 85–91.
16. R. Kumar, S. Mathias, S. Saha and P. Bhattacharyya, *Many hands make light work: Using essay traits to automatically score essays*, arXiv preprint, 2021, DOI 10.48550/arXiv.2102.00781
17. V. Kumar and D. Boulanger, *Explainable automated essay scoring: Deep learning really has pedagogical value*, Front. Educ. **5** (2020), DOI 10.3389/feduc.2020.572367
18. H. Manabe and M. Hagiwara, *EXPATS: a toolkit for explainable automated text scoring*, *arXiv preprint*, 2021, DOI 10.48550/arXiv.2104.03364
19. S. Mathias and P. Bhattacharyya, *ASAP++: Enriching the ASAP automated essay grading dataset with essay attribute scores*, (Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018), Miyazaki, Japan), 2018.
20. ETS Corpus of Non-Native Written English, (2014). https://catalog.ldc.upenn.edu/LDC2014T06
21. S. Prabhu, K. Akhila and S. Sanriya, *A hybrid approach towards automated essay evaluation based on BERT and feature engineering*. (IEEE 7th International conference for Convergence in Technology (I2CT), Mumbai, India), 2022, DOI 10.1109/I2CT54291.2022.9824999
22. M. Chen and X. Li, *Relevance-based automated essay scoring via hierarchical recurrent model*. (International Conference on Asian Language Processing (IALP), Bandung, Indonesia), 2018, pp. 378–383.
23. H. Bai, Z. Huang, A. Hao, and S. C. Hui, *Gated character-aware convolutional neural network for effective automated essay scoring*, (IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, Melbourne, Australia), 2021, pp. 351–359.
24. J. Schneider, R. Richner, and M. Riser, *Towards trustworthy autograding of short, multi-lingual, multi-type answers*, Int. J. Artif. Intell. Educ. **33** (2023), no. 1, 88–118.
25. T. Mizumoto, H. Ouchi, Y. Isobe, P. Reisert, R. Nagata, S. Sekine and K. Inui, *Analytic score prediction and justification identification in automated short answer scoring*, (Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications, Florence, Italy), 2019, pp. 316–325.
26. Y. Wang, C. Wang, R. Li and H. Lin, *On the use of bert for automated essay scoring: Joint learning of multi-scale essay representation*, arXiv preprint, 2022, DOI 10.48550/arXiv.2205.03835
27. C. M. Ormerod, A. Malhotra and A. Jafari, *Automated essay scoring using efficient Transformer-based language models*, arXiv preprint, 2021, DOI 10.48550/arXiv.2102.13136
28. A. Mizumoto and M. Eguchi, *Exploring the potential of using an AI language model for automated essay scoring*, Res. Method Appl. Linguist. **2** (2023), no. 2, DOI 10.1016/j.rmal.2023.100050
29. X. Li, M. Chen, and J. Y. Nie, *SEDNN:* shared and enhanced deep neural network model for cross-prompt automated essay scoring, Knowledge-Based Syst. **210** (2020), DOI 10.1016/j.knosys.2020.106491
30. R. Ridley, L. He, X. Dai, S. Huang and J. Chen, *Prompt agnostic essay scorer: a domain generalization approach to cross-prompt automated essay scoring*, arXiv preprint, 2020, DOI 10.48550/arXiv.2008.01441
31. Y. Cao, H. Jin, X. Wan and Z. Yu, *Domain-adaptive neural automated essay scoring*, (Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, China), 2020, pp. 1011–1020.
32. C. Jin, B. He, K. Hui and L. Sun, *TDNN: a two-stage deep neural network for prompt-independent automated essay scoring*, (Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, Melbourne, Australia), 2018, pp. 1088–1097.
33. H. Funayama, Y. Asazuma, Y. Matsubayashi and T. M. K. Inui, *What can short answer scoring models learn from cross-prompt training data?* (Language Processing Society 29th Annual Conference (NLP2023), Okinawa), 2023, pp. 1874–1879.
34. R. Ridley, L. He, X. Y. Dai, S. Huang, and J. Chen, *Automated cross-prompt scoring of essay traits*, Proc AAAI Conf. Artif. Intell. **35** (2021), no. 15, 13745–13753.
35. X. Wang, Y. Lee and J. Park, *Automated evaluation for student argumentative writing: A survey*, arXiv preprint, 2022, DOI 10.48550/arXiv.2205.04083
36. Y. He, F. Jiang, X. Chu and P. Li, *Automated Chinese Essay Scoring from Multiple Traits*, (Proceedings of the 29th International Conference on Computational Linguistics, Gyeongju, Rep. of Korea), 2022, pp. 3007–3016.
37. W. Song, Z. Song, L. Liu and R. Fu, *Hierarchical multi-task learning for organization evaluation of argumentative student essays*, (Proc. International Joint Conference on Artificial Intelligent), 2021, pp. 3875–3881.
38. D. Liao, J. Xu, G. Li, and Y. Wang, *Hierarchical coherence modeling for document quality assessment*, Proc. AAAI Conf. Artif. Intell. **35** (2021), no. 15, 13353–13361.
39. F. S. Mim, N. Inoue, P. Reisert, H. Ouchi and K. Inui, *Unsupervised learning of discourse-aware text representation for essay scoring*, (Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop, Florence, Italy), 2019, pp. 378–385.
40. T. Abhishek, D. Rawat, M. Gupta and V. Varma, *Transformer models for text coherence assessment*, arXiv preprint, 2021, DOI 10.48550/arXiv.2109.02176

41. S. Behzad, A. Zeldes and N. Schneider, *Sentence-level Feedback Generation for English Language Learners: Does Data Augmentation Help?* arXiv preprint, 2022, DOI 10.48550/arXiv.2212.08999
42. R. Nagata, *Toward a task of feedback comment generation for writing learning*, (Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China), 2019, pp. 3206–3215.
43. K. Hanawa, R. Nagata and K. Inui, *Exploring methods for generating feedback comments for writing learning*, (Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing), 2021, pp. 9719–9730.
44. R. Nagata, M. Hagiwara, K. Hanawa, M. Mita, A. Chernodub and O. Nahorna, *Shared task on feedback comment generation for language learners*, (Proceedings of the 14th International Conference on Natural Language Generation, Aberdeen, Scotland), 2021, pp. 320–324.
45. S. Coyne, *Template-guided Grammatical Error Feedback Comment Generation*, (Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop, Dubrovnik, Crotia), 2023, pp. 94–104.
46. Z. Zhang, J. Guan, G. Xu, Y. Tian and M. Huang, *Automatic Comment Generation for Chinese Student Narrative Essays*, (Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Abu Dhabi, UAE), 2022, pp. 214–223.
47. F. Dong, Y. Zhang and J. Yang, *Attention-based recurrent convolutional neural network for automatic essay scoring*, (Proceedings of the 21st conference on computational natural language learning, Vancouver, Canada), 2017, pp. 153–162.
48. S. Jeon and M. Strube, *Countering the influence of essay length in neural essay scoring*, (Proceedings of the Second Workshop on Simple and Efficient Natural Language Processing), 2021, pp. 32–38.
49. M. Zaheer, G. Guruganesh, K. A. Dubey, J. Ainslie, C. Alberti, S. Ontanon, P. Pham, A. Ravula, Q. Wang, L. Yang, and A. Ahmed, *Big bird: transformers for longer sequences*, Adv. Neur. Inf. Process Syst. **33** (2020), 17283–17297.
50. https://www.kaggle.com/competitions/asap-aes/data
51. https://www.kaggle.com/code/javigallego/english-language-learning-complete-eda
52. J. Cohen, *Weighted kappa: nominal scale agreement provision for scaled disagreement or partial credit*, Psychol. Bull. **70** (1968), no. 4, 213–220.

AUTHOR BIOGRAPHIES



Minsoo Cho received a B.S. degree in Information Communication Engineering from Dongguk University, Seoul, Republic of Korea, in 2017 and an M.S. degree in Computer Science Engineering from Yonsei University, Seoul, Republic of Korea, in 2019. Since 2020, she has been working with the Language Intelligence Research Section, Electronics and Telecommunications Research Institute, Daejeon, Republic of Korea, as a researcher. Her research interests include natural language processing, intelligent tutoring, and dialogue systems.

Jin-Xia Huang received a B.S. degree in physics from Jilin University, China, in 1991, an M.S. degree in computer science from KAIST, Republic of Korea, in 2001, and a Ph.D. degree in computer science from Jeonbuk National University, Republic of Korea, 2018. From 2001 to 2003, she worked as a Researcher for Microsoft Research Asia in Beijing, China. Since 2008, she has been working at the Language Intelligent Research Section of the Electronics and Telecommunications Research Institute (ETRI), Daejeon, Korea, where she is currently a Principal Researcher. Her research interests focus on natural language processing, especially dialogue systems and intelligent tutoring at present.

Oh-Woog Kwon received the B.S. degree in computer engineering from Kyungpook National University, Republic of Korea, in 1992, the M.S. degree in computer science from KAIST, Republic of South Korea, in 1995, and the Ph.D. degree in computer engineering from the Pohang University of Science and Technology (POSTECH), Republic of Korea, in 2001. Since 2004, he has been working with the Language Intelligent Research Section, Electronics and Telecommunications Research Institute (ETRI), Daejeon, Republic of Korea, where he is currently a Director. His major research interests include natural language processing, large language models, and mathematical reasoning.

How to cite this article: M. Cho, J.-X. Huang, and O.-W. Kwon, *Dual-scale BERT using multi-trait representations for holistic and trait-specific essay grading*, ETRI Journal **46** (2024), 82–95, DOI 10.4218/etrij.2023-0324.

APPENDIX A

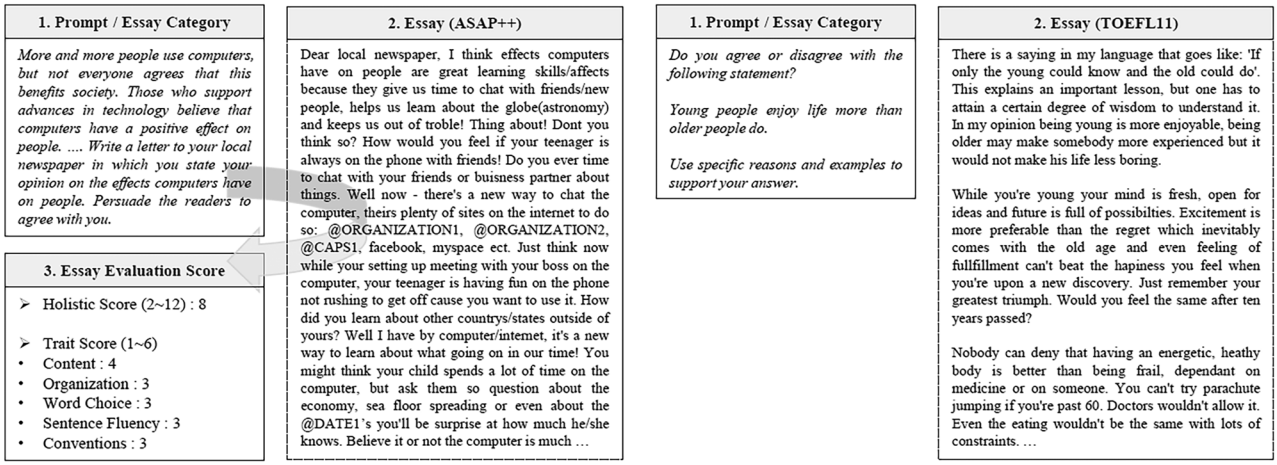


FIGURE A1 Examples of an essay and prompt from both ASAP++ and TOEFL11 dataset.