Bulletin of Electrical Engineering and Informatics
Vol. 15, No. 3, June 2026, pp. 2327~2344
ISSN: 2302-9285, DOI: 10.11591/eei.v15i3.11235

A BERT-based modular framework for automated English essay scoring via trait analysis

Jasman Pardede, Rizka Milandga Milenio, **Thalita Zharifa Nathania**
Department of Informatics, Faculty of Industrial Technology, Institut Teknologi Nasional Bandung, Bandung, Indonesia

Article Info

Article history:

Received Aug 19, 2025
Revised Feb 13, 2026
Accepted Mar 10, 2026

Keywords:

Automated essay scoring
Bidirectional encoder representations from transformer
Grammar scoring
Structure scoring
Trait-based evaluation

ABSTRACT

Automated essay scoring (AES) systems are commonly implemented using holistic scoring, which limits interpretability and prevents assessment at the writing trait level. As a result, such systems provide limited diagnostic and actionable feedback. To address this limitation, this study proposes a modular trait-based AES framework that separates structure and grammar evaluation while maintaining an integrated scoring mechanism. The proposed framework consists of two modules. The structure module evaluates the ideas, organization, and style traits using a bidirectional encoder representations from transformer-bidirectional long short-term memory (BERT-BiLSTM-Attention) architecture trained on the automated student assessment prize (ASAP) dataset. The grammar module evaluates the Conventions trait by applying a BERT-based grammatical acceptability classifier trained on the Corpus of linguistic acceptability (CoLA) dataset, followed by multinomial logistic regression to convert grammatical patterns into interpretable grammar scores. Experiments were conducted on the ASAP dataset and evaluated using the quadratic weighted Kappa (QWK) metric. The structure module achieved a QWK score of 0.7906 on the test set, while the grammar module obtained a QWK of 0.3923. The integrated holistic score reached a QWK of 0.7847. These results demonstrate that the proposed modular framework improves interpretability and scoring performance, supporting more objective and actionable essay evaluation for formative assessment in English language education.

Corresponding Author:

Jasman Pardede
Department of Informatics, Faculty of Industrial Technology, Institut Teknologi Nasional Bandung
St. PHH Mustofa No. 23, Cikutra, Cibeunying Kidul, Bandung 40124, West Java, Indonesia
Email: jasman@itenas.ac.id

1. INTRODUCTION

Essay assessment plays a crucial role in educational evaluation as it reflects students' critical thinking, reasoning ability, and written communication skills. In most educational settings, essay scoring is traditionally performed manually using holistic judgment, where evaluators assign a single overall score based on their subjective impression of the essay quality. Although this approach is widely adopted due to its simplicity, it suffers from several inherent limitations, including subjectivity, inter-rater inconsistency, and evaluator fatigue, all of which may compromise fairness and reliability in large-scale assessments [1].

To address these issues, automated essay scoring (AES) systems have been extensively studied as an alternative that leverages natural language processing (NLP) and machine learning (ML) techniques to automatically evaluate student essays [2]. Early AES systems predominantly relied on surface-level features such as essay length, word frequency, vocabulary richness, and spelling or grammar errors. While computationally efficient, these features fail to capture deeper semantic meaning and discourse coherence,

allowing essays with weak arguments but mechanically correct writing to receive disproportionately high scores [3]-[5]. Several publicly available datasets have been introduced to support AES research, including the automated student assessment prize (ASAP) dataset, which provides trait-level annotations for essay attributes [6]. These datasets have facilitated the development of more fine-grained scoring models and highlighted the importance of moving beyond surface-level features toward deeper semantic and structural analysis.

Recent advances in deep learning have significantly improved AES performance through the adoption of neural architectures such as convolutional neural networks (CNN), long short-term memory (LSTM) networks, and Transformer-based models. Among these approaches, bidirectional encoder representations from transformers (BERT) has demonstrated strong capability in modeling contextual and semantic relationships within text, leading to notable performance gains in AES tasks [7]-[9]. Subsequent studies further show that fine-tuning BERT for AES enables task-specific adaptation and significantly improves scoring accuracy compared to using frozen embeddings [10], [11].

Despite these improvements, most existing BERT-based AES systems still adopt a holistic scoring paradigm, producing a single overall score for each essay. This design limits interpretability and fails to provide diagnostic feedback on specific writing traits such as ideas, organization, style, and grammatical correctness. Some studies attempt to incorporate grammar-related features into BERT-based models; however, grammar is typically treated as an auxiliary feature embedded within a single scoring model rather than as an independently evaluated trait [7], [12]. In parallel, research on grammatical acceptability has shown that sentence-level grammar judgments can be effectively modeled using neural classifiers trained on the Corpus of linguistic acceptability (CoLA) dataset, providing a principled foundation for explicit grammar evaluation in downstream applications [13].

Recent studies in explainable AES emphasize the growing need for transparent and diagnostically meaningful assessment systems. Beyond predictive performance, educational applications require models that provide interpretable feedback aligned with specific writing traits. This demand reinforces the importance of modular and trait-oriented architectures that can deliver both reliable scoring and pedagogically actionable insights [14].

Recent large language model (LLM)-based AES studies further demonstrate that although transformer architectures significantly improve semantic modeling, many systems still rely on unified scoring mechanisms that limit interpretability. Hybrid frameworks integrating contextual embeddings with explicit trait modeling show improved robustness and diagnostic capability, highlighting the importance of modular and interpretable scoring designs in educational assessment [15].

To address this gap, this paper proposes a novel modular trait-based AES framework that explicitly decomposes essay assessment into distinct yet complementary components. The proposed system consists of two specialized modules. The structure module evaluates higher-level writing traits ideas, organization, and style using a bidirectional encoder representations from transformer-bidirectional long short-term memory (BERT-BiLSTM-Attention) architecture trained on the ASAP dataset. This module focuses on semantic coherence, discourse structure, and stylistic consistency. In parallel, the grammar module evaluates the Conventions trait using a BERT-based grammatical acceptability classifier trained on the CoLA dataset, followed by multinomial logistic regression to map detected grammatical patterns into interpretable grammar scores.

To evaluate the effectiveness of the proposed framework, quadratic weighted Kappa (QWK) is employed as the primary performance metric, as it is widely adopted in AES research due to its suitability for ordinal scoring and its ability to penalize larger prediction errors more severely. Recent studies emphasize the importance of QWK in ensuring fair and reliable comparison between automated and human essay scores [16]. In addition to evaluation considerations, prior work on transformer-based AES in non-English contexts demonstrates the robustness and adaptability of contextual language models across different languages and educational settings, supporting the generalizability of the proposed approach [17].

The key novelty of this work lies in its explicit modularization of grammar assessment as a standalone classification-based component, which fundamentally differs from prior BERT-based AES approaches that embed grammar features within a single holistic scoring model. By decoupling semantic structure evaluation and grammatical assessment into independent yet integrated modules, the proposed framework enhances trait-level interpretability, enables more precise diagnostic feedback, and improves computational efficiency while maintaining competitive scoring performance. Unlike most existing AES approaches, this study integrates a BERT-based semantic scoring module and a BERT-based grammar evaluation module within a unified yet modular trait-based framework.

Unlike prior BERT-based AES systems that rely on holistic or unified multi-trait prediction within a single architecture, the proposed framework adopts an explicitly modular design that separates semantic structure evaluation from grammatical competence assessment. Previous approaches typically embed

grammar implicitly within shared essay representations. In contrast, this study trains two independently optimized modules using distinct datasets (ASAP for structure and CoLA for grammar), enabling task-specific learning and reducing representational interference. Furthermore, grammar scores are derived from interpretable linguistic features extracted from a dedicated acceptability classifier, enhancing diagnostic transparency and improving flexibility compared to monolithic BERT-based scoring models.

The main contributions of this study are threefold. First, this work proposes a modular trait-based AES architecture that explicitly separates structure and grammar evaluation while preserving an integrated scoring framework, enabling more focused and interpretable assessment of distinct writing aspects. Second, a hybrid grammar scoring mechanism is introduced by combining BERT-based grammatical acceptability detection with statistical classification, allowing grammatical errors to be translated into transparent and diagnostically meaningful scores. Third, this study demonstrates that trait-level scoring provides significant advantages over conventional holistic AES systems by improving interpretability, fairness, and the system's ability to diagnose specific strengths and weaknesses in student writing. The proposed framework advances AES beyond single-score prediction by providing more transparent, objective, and actionable feedback, making it particularly suitable for formative assessment and educational applications that aim to support writing development.

2. METHOD

Explaining this research was trained using two main types of datasets, namely ASAP dataset and CoLA dataset. The ASAP is used as the main source for all stages of the system because it contains a collection of student essays that have been scored by human raters, making it highly relevant for training and testing AES. Meanwhile, the CoLA is used specifically to train the grammar checking module, as CoLA provides grammaticality labels for thousands of sentences, allowing the model to learn to distinguish grammatical and ungrammatical sentences.

To further clarify the procedural flow of the proposed system, Algorithm 1 presents the pseudocode of the modular trait-based AES pipeline from input preprocessing to holistic score generation.

Algorithm 1. Modular trait-based AES pipeline
Input:
Preprocessed essays from the ASAP dataset
Output:
Trait-level scores (Ideas, Organization, Style, Conventions) and holistic score
Steps:

1. Tokenize each essay using a BERT-compatible tokenizer.
2. Encode the essay using pretrained BERT embeddings.
3. Pass embeddings through a BiLSTM layer followed by attention pooling.
4. Predict Ideas, Organization, and Style scores.
5. Compute the structure score using a weighted aggregation of the three traits.
6. Split the essay into sentences for grammar analysis.
7. Classify grammatical acceptability using a BERT-based grammar checker.
8. Extract grammar-related features from sentence-level predictions.
9. Predict the grammar score using multinomial logistic regression.
10. Combine structure and grammar scores to produce the holistic score.

Figure 1 illustrates the overall workflow of the proposed modular AES system, including data preprocessing, independent training of structure and grammar modules, and the final holistic score aggregation. The proposed system comprises three main model training stages: i) structure scoring, ii) grammar checking, and iii) grammar scoring. In the first stage, a structure scoring model is trained to assess three key traits of essay structure ideas, organization, and style [9]. The model leverages BERT embeddings, which are passed through a bidirectional long short-term memory (BiLSTM) network followed by an attention layer to capture contextual relationships across essay sentences. Essays are tokenized into BERT-compatible input and trained using structure trait annotations from the ASAP dataset. The model outputs a score for each trait and a total structure score calculated as a weighted combination of the three.

The second stage involves training a BERT-based binary classifier using the CoLA dataset to detect grammatical acceptability at the sentence level. This grammar checking model is later used to evaluate grammatical correctness across essay sentences. The third stage constructs a grammar scoring model using logistic regression. Numerical features are derived from the grammar checking outputs, such as the number of ungrammatical sentences per essay and per clause. These features are fed into a logistic regression model trained on grammar-annotated ASAP essays to predict grammar scores.

Upon completing all training stages, the system produces two primary outputs: a structure score (comprising three traits) and a grammar score. These are then combined into a holistic score based on predetermined weighting. Model performance covering structure, grammar, and holistic scoring is evaluated using the QWK metric. QWK is suitable for ordinal scoring tasks, as it accounts for both agreement and magnitude of error between predicted and human-assigned scores [16], [18], [19], making it widely adopted in essay scoring systems.

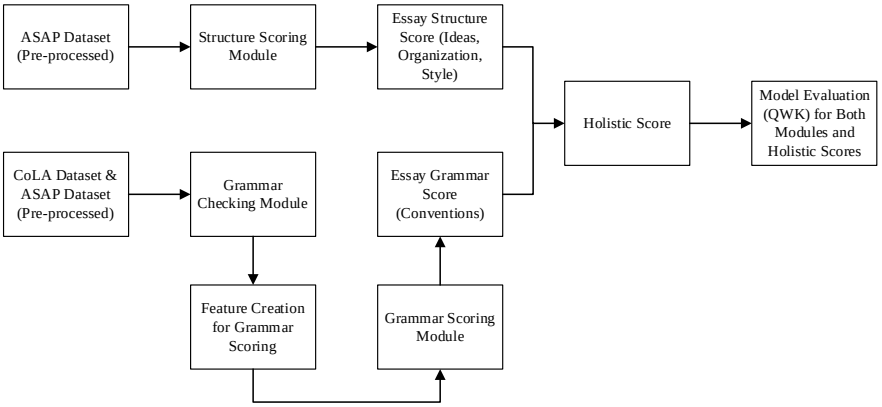

Figure 1. Workflow of the proposed modular trait-based AES system, including preprocessing, module training, and score aggregation

2.1. Dataset

This research uses two complementary datasets: the ASAP AES dataset and the CoLA dataset, selected to support the two core components of the AES system-structure and grammar assessment [20], [21]. The ASAP dataset, sourced from a public Kaggle competition by the Hewlett Foundation, is used to train and evaluate the model in scoring essay structure. Among its eight essay sets, this study focuses on Set 7, written by 7th-grade students, as it includes trait-level annotations for ideas, organization, style, and grammar [6]. The dataset contains 1,569 essays, each averaging 11 lines and 250 words, and scored by two raters. The essays cover narrative, persuasive, and expository genres. Table 1 presents the raw data structure of the ASAP dataset prior to preprocessing.

Table 1. Raw entries from the ASAP AES dataset with essay IDs, set information, excerpts, and trait-level scores from two raters (R1 and R2)

Essay_id	Essay_set	Essay	R1_T1	R1_T2	R1_T3	R1_T4	R2_T1	R2_T2	R2_T3	R2_T4
17834	7	Patience is when your waiting. I was pa...	1	2	2	3	1	2	2	2
17836	7	I am not a patience person, like I can't...	1	1	2	2	2	2	2	1
17837	7	One day I was at basketball practice and...	1	2	2	2	2	2	2	2
17838	7	I going to write about a time when I was...	2	2	2	2	2	2	2	3

The ASAP dataset includes the essay ID, essay set, and full essay text as model input. Each essay is rated by two annotators on four traits: ideas, organization, style, and conventions, each ranging from 0-3. The ideas score is multiplied by two, resulting in a total score between 0-15 based on the official rubric.

The second dataset, the CoLA, is used to train the grammar checking model. It contains 10,657 expert-annotated sentences labeled for grammatical acceptability, collected from 23 linguistic publications [22]. A BERT-based binary classifier is trained on CoLA and applied to sentences in the ASAP essays. The outputs are then aggregated into numerical features for grammar score prediction. Table 2 illustrates the sentence-level structure of the CoLA dataset used for grammatical acceptability classification.

The CoLA dataset has four columns: a source code indicating the origin of the sentence, a binary label for grammatical acceptability (1=acceptable, 0=unacceptable), the original author's acceptability rating, and the sentence itself. Together, ASAP and CoLA provide a complementary foundation for the AES system: ASAP supports scoring of essay content and structure, while CoLA enables grammatical error detection.

Table 2. CoLA dataset format with source, acceptability labels (1=acceptable, 0=unacceptable), linguistic judgment labels, and sentence text

Source code	Label	OrigLabel	Sentence
gj04	1		The sailors rode the breeze clear of the rocks.
gj04	1		The weights made the rope stretch over the pulley.
gj04	1		The mechanical doll wriggled itself loose.
cj99	1		If you had eaten more, you would want less.
cj99	0	*	As you eat the most, you want the least.

2.2. Data pre-processing

The data pre-processing procedure began with filtering the ASAP dataset to include only essay set 7, ensuring rubric consistency for multi-trait scoring. Each essay was scored by two raters; the average score per trait was used to reduce inter-rater variability, then rounded to a 0-3 scale following the official rubric and prior research [10]. Structure assessment focused on three traits ideas, organization, and style distinct from the grammar trait (conventions), in line with standard essay evaluation practices [9]. The final structure score was calculated using the formula: 2×Ideas+Organization+Style, emphasizing idea development [9]. The grammar score was taken directly from the Conventions trait using the same averaging method.

The cleaned dataset, consisting of essay text, trait scores, structure score, and grammar score, is presented in Table 3. Irrelevant features were removed, retaining only essential columns such as essay ID, text, and scoring information. After preprocessing, the dataset (1,569 essays) was split into training (70%), validation (15%), and testing (15%) sets using the train_test_split function. This division ensures balanced training, helps prevent overfitting, and enables reliable evaluation on unseen data.

The distribution of the number of essays in each set can be seen in Table 4, which illustrates the data split to effectively support the training, validation, and testing processes of the model. This pre-processing process is essential to ensure that the data is in the right format and contains the features needed to effectively and accurately train and evaluate the AES system.

Table 3. Preprocessed ASAP dataset entries with averaged rater scores and weighted structure score (2×Ideas+Organization+Style)

Essay_id	Essay (excerpt)	Ideas_score	Organization_score	Style_score	Grammar_score	Structure_true_score
18776	A time when someone else I knew...	3	3	2	3	11
17984	Patience is a virtue. It's something...	2	2	2	3	8
19215	I had a friend well she is @CAPS4...	3	3	2	2	11
18471	Have you ever have to wait in line...	2	2	2	3	8

Table 4. Distribution of essays across training, validation, and testing sets after dataset splitting with proportions of 70%, 15%, and 15%, respectively

Dataset	Number of essays	Percentage (%)
Training set	1098	70
Validation set	235	15
Testing set	236	15

2.3. Structure scoring module

As shown in Figure 2, the module begins by tokenizing the essay using the BERT tokenizer (bert-base-uncased), converting text into token IDs and attention masks. Since BERT limits input to 512 tokens, a sliding window strategy is used to handle longer essays [23]. Following Prabhu's method, the text is split into overlapping segments (e.g., [0-512] and [256-768]) to retain context [7]. If tokens are fewer than 512, [PAD] tokens are added, and the attention mask assigns 1 to real tokens and 0 to padding.

Each chunk is passed to the 12-layer BERT model, where multi-head self-attention and feed-forward layers produce 768-dimensional contextual embeddings. These embeddings are fed into a two-layer LSTM: a larger first layer for capturing sequence information and a smaller second layer to reduce complexity. Dropout is applied to prevent overfitting. Then, attention pooling is used to weigh token importance and ignore padding during aggregation [12]. More relevant tokens are given more weight, while less important tokens, including padding, are ignored in the aggregation process.

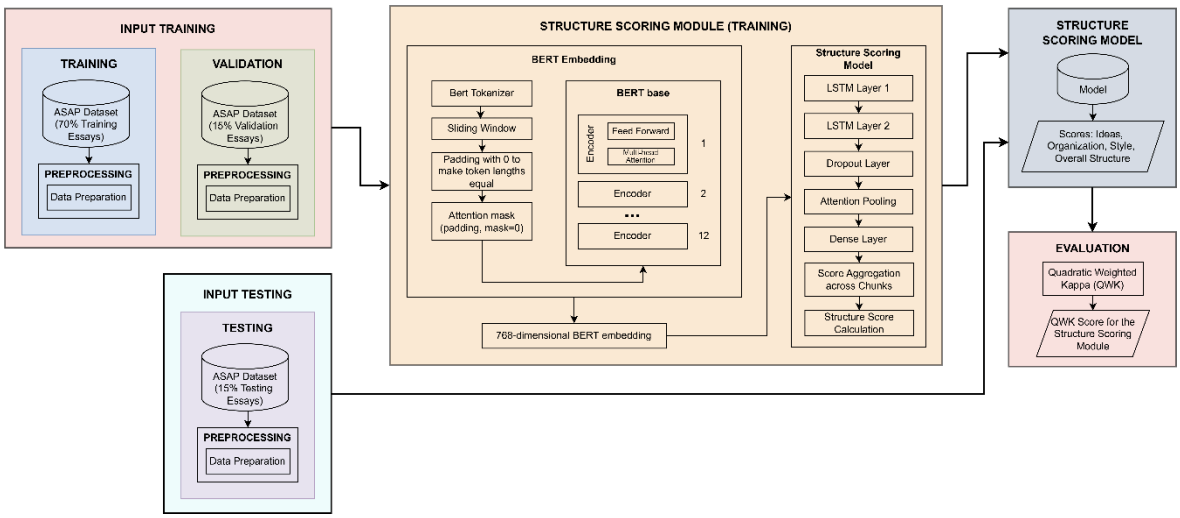

Figure 2. BPM of the structure scoring module with tokenization, BERT embedding, BiLSTM-attention, and trait score prediction

Recent research combining contextual sentence embeddings with LSTM-Attention architectures demonstrates that attention mechanisms significantly enhance the model's ability to focus on salient textual components. This selective weighting strategy improves representation quality and contributes to better generalization in essay scoring tasks [24].

Each chunk is transformed into a fixed-size global vector representing its contextual content, which is then passed to three separate fully connected layers to predict ideas, organization, and style scores. These use ReLu activation and output float regression values, later rounded for evaluation. If multiple chunks exist, trait scores are averaged. The final structure score is computed as: 2×Ideas+Organization+Style.

This module outputs both individual trait scores and the overall structure score based on rubric weighting. It follows a multi-trait scoring approach for more accurate and aspect-specific evaluation [9], [12], [25], with performance assessed using the QWK metric.

During training, the structure scoring model was fine-tuned using the Adam optimizer with a learning rate of 2e-5, following common practice for Transformer-based models. The batch size was set to 16 to balance memory constraints and training stability. The model was trained for 50 epochs, with early stopping based on validation loss to mitigate overfitting. Dropout regularization was applied within the BiLSTM layers to improve generalization. These hyperparameter values were selected based on preliminary experiments and prior studies employing BERT for essay scoring tasks.

2.4. Grammar checking model

As shown in Figure 3, the process starts by tokenizing each sentence using the bert-base-uncased tokenizer, which converts text into token IDs and attention masks. Token IDs represent words/subwords numerically, while the attention mask marks real tokens with 1 and [PAD] tokens with 0. All inputs are standardized to 512 tokens; shorter sentences are padded accordingly. The attention mask ensures padding tokens are ignored during BERT's attention mechanism.

The tokenized inputs are then passed to the BERT base model, which has 12 encoder layers with multi-head self-attention and feed-forward networks. BERT generates contextual embeddings for each token, and the [CLS] vector is used as a global representation of the sentence for classification tasks [26]-[28]. This [CLS] output is fed into a binary classifier to predict whether a sentence is grammatical (label 1) or ungrammatical (label 0).

Training uses HuggingFace's trainer, and evaluation metrics include accuracy, precision, recall, F1-score, and confusion matrix. The resulting model is a BERT-based grammatical classifier, stored for use in the next stage: predicting grammaticality in ASAP essay sentences. These predictions are converted into numerical features for input into the grammar scoring regression module.

The grammar checking model was fine-tuned on the CoLA dataset using the bert-base-uncased architecture. Training employed the AdamW optimizer with a learning rate of 2e-5 and a batch size of 16. The model was trained for 4 epochs, which was sufficient to achieve stable convergence on validation performance. Padding and attention masking were applied to standardize input length to 512 tokens. These settings align with standard configurations for sentence-level grammatical acceptability classification.

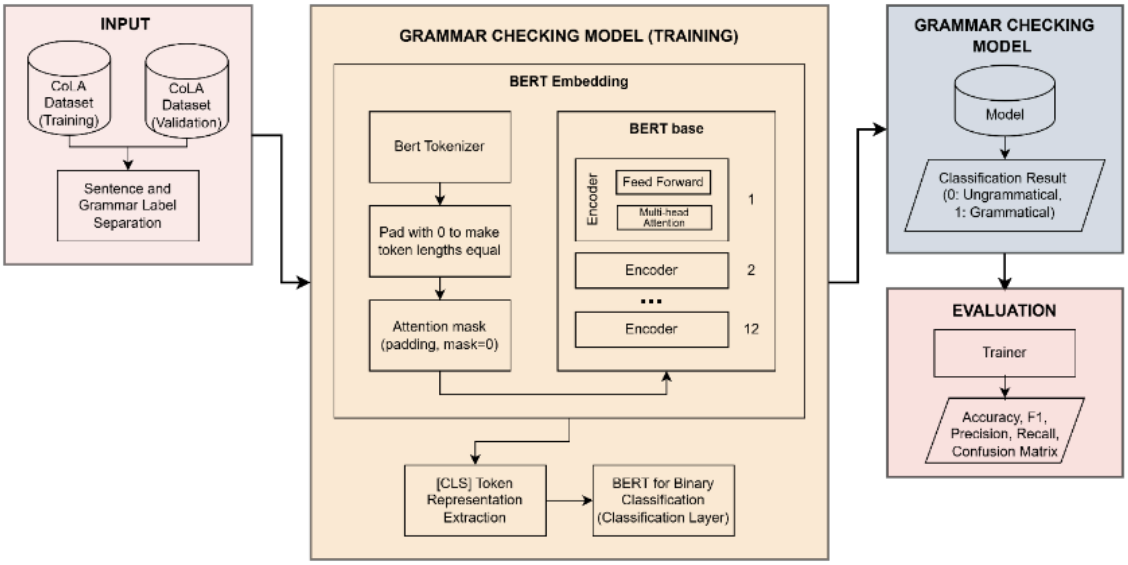

Figure 3. BPM of the grammar checking module with sentence tokenization, BERT encoding, and binary acceptability classification using the [CLS] token

2.5. Grammar scoring module

The grammar scoring module in the AES system evaluates essay grammar quality through two main components: a BERT-based grammar checker and a logistic regression score mapping model. The grammar checker classifies each sentence and clause as “grammatical” (1) or "ungrammatical" (0), while the second component maps these outputs into a final grammar score.

As illustrated in Figure 4, the process begins with preprocessed essays from the ASAP dataset. Sentences and clauses are extracted using the spaCy library for syntactic parsing. Each extracted unit is then passed into a grammar checker model fine-tuned on the CoLA dataset, using the [CLS] token for binary classification.

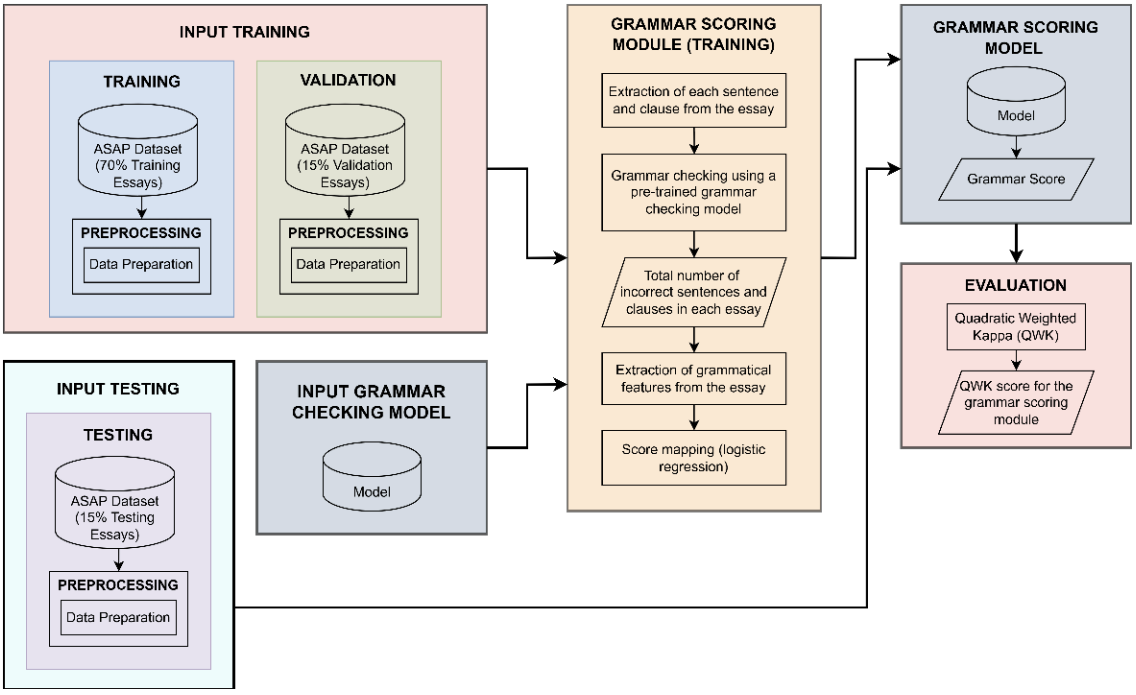

Figure 4. BPM of the grammar scoring module with prediction aggregation and multinomial logistic regression

The classification results are aggregated to calculate counts of grammatical units and errors per essay. Based on Tambe's approach [1], these statistics are converted into numerical features, including sentence count, clause count, number of errors, and spelling correction outputs. These features serve as input to a multinomial logistic regression model [29], [30].

Before training, features are standardized using StandardScaler to ensure uniform scaling. The logistic regression model, using SoftMax for multi-class prediction, maps feature combinations to grammar scores ranging from 0-3. Training uses gradient descent to minimize loss until convergence [29]. The resulting model predicts grammar scores aligned with human ratings and is evaluated using the QWK metric.

The grammar scoring model utilized a multinomial logistic regression classifier trained on grammar-related features extracted from the grammar checking stage. The model was optimized using gradient descent with a maximum iteration limit of 2000 to ensure convergence. Prior to training, all features were standardized using z-score normalization. No additional regularization was applied, as the feature dimensionality was limited and empirically observed to be stable during training.

2.6. Holistic scoring and weighting scheme

After obtaining trait-level predictions from the structure and grammar modules, a weighted aggregation scheme is applied to compute the holistic score. The structure score is calculated as a weighted combination of ideas, organization, and style, where the ideas trait receives double weight to reflect its higher importance in essay quality assessment. The final holistic score is computed by summing the structure score and the grammar score. This weighting scheme follows the scoring rubric of the ASAP dataset and is consistently applied to both reference scores and model predictions.

The first step in the calculation is to determine the structure score, which consists of three main traits. These three are predicted separately by the BERT and BiLSTM-based models, and then combined using the weighting formula set out in the scoring rubric, as shown in (1):

$$Structure\ Score = 2 \times Idea + Organization + Style \quad (1)$$

Giving twice the weight to ideas refers to the rubric used in the ASAP competition and is supported by Attali and Sinharay (2015) [9], who emphasize the importance of idea development in assessing essay structure. The holistic score aggregation follows an explicit weighting scheme in which content-related traits receive higher emphasis than language conventions, in accordance with the ASAP scoring rubric. The calculation process is done by summing up the two main components, namely the structure score as well as the grammar score [1]. The merging of the two scores is mathematically formulated as shown in (2):

$$\text{Holistic Score} = \text{Structure Score} + \text{Grammar Score} \quad (2)$$

This calculation process is applied both for the reference score of the dataset (true score) and for the model prediction result (pred score), so that the model performance evaluation can be carried out consistently.

2.7. Model evaluation method

The performance of the models in this AES system was evaluated using the QWK metric. QWK was selected as it effectively quantifies agreement between two raters in this case, between the model's predicted scores and human-assigned reference scores while accounting for the ordinal nature of the data [16], [18]. Unlike simple accuracy, QWK penalizes predictions that deviate further from the true score. This makes QWK well-suited for essay scoring, where the degree of disagreement matters.

Recent AES studies reaffirm that QWK remains the preferred evaluation metric for essay scoring systems due to its suitability for ordinal prediction tasks and its sensitivity to the magnitude of disagreement between predicted and human scores [31]. The mathematical formulation of QWK is presented in (3), following the approach described by Chilukoti *et al.* [18]:

$$QWK = 1 - \frac{\Sigma_{ij} W_{ij} O_{ij}}{\Sigma_{ij} W_{ij} E_{ij}} \quad (3)$$

QWK is a metric that measures the agreement between two raters in this case, the model and a human evaluator while accounting for the ordinal nature of the scores. Here, *i* denotes the actual label from the human rater, and j is the model's predicted label. The term O_{ij} refers to the observed count of times label *i* was paired with prediction *j*, forming the observation (confusion) matrix. E_{ij} represents the expected count for that pair if predictions were random, forming the expectation matrix. Finally, W_{ij} is the weight of disagreement between labels *i* and *j*, calculated using the quadratic weighting formula as shown in (4):

$$W_{ij} = \frac{(i-j)^2}{(N-1)^2} \quad (4)$$

where N is the number of score classes available. QWK values are in the range -1 to 1 [19]. Model evaluation is performed separately on three data scenarios: training, validation, and testing. QWK was calculated for each of the assessed aspects (structure, grammar), as well as for the holistic score.

Model evaluation followed a fixed data split strategy, where the dataset was partitioned into training, validation, and testing sets with proportions of 70%, 15%, and 15%, respectively. Cross-validation was not employed, as the ASAP dataset provides predefined essay sets and the computational cost of repeated training for transformer-based models is substantial. Random seed initialization was not explicitly fixed during model training; however, all experiments adhered to the same data partitioning scheme and evaluation protocol to ensure consistency and comparability across experimental results.

3. RESULTS AND DISCUSSION

3.1. Structure module scoring

Early experiments using a holistic scoring approach demonstrated strong overall agreement; however, the single-score output limited interpretability and did not provide diagnostic feedback. When the model was extended to predict individual structure traits, the output became more informative, but performance degradation indicated that the pooling mechanism remained a critical limitation.

To better capture key information, mean pooling was replaced with attention pooling, enabling the model to focus on more relevant tokens. Replacing mean pooling with attention pooling enables the model to selectively emphasize salient textual elements, resulting in more effective essay representations and improved generalization [12].

As summarized in Table 5, replacing mean pooling with attention pooling consistently improves validation performance, indicating that the model benefits from selectively focusing on salient textual features rather than treating all tokens equally.

Table 6 indicates that a moderate learning rate provides the best balance between convergence speed and generalization. Extremely small learning rates result in slower learning, while larger values tend to reduce validation stability. The learning rate analysis shows that moderate values provide a better balance between convergence stability and generalization. Extremely small learning rates slow optimization, while larger values increase sensitivity to validation fluctuations. Based on this trend, a learning rate of 2e-5 was selected for subsequent experiments.

Table 5. Validation QWK comparison between holistic and trait-based structure scoring approaches using mean pooling and attention pooling strategies

Approach	Output scheme	Pooling	QWK training	QWK validation
Holistic	1 total score	Mean pooling	0.9976	0.8253
Trait-based	3 traits+total	Mean pooling	0.9997	0.7673
Trait-based	3 traits+total	Attention	0.9997	0.8022

Table 6. Effect of different learning rates on structure module performance measured using QWK and final training loss

Learning rate	QWK training	QWK validation	Final loss
1e-5	0.9992	0.7700	0.0153
2e-5	0.9997	0.8022	0.0036
5e-6	0.9985	0.7790	0.0223

After selecting the optimal learning rate, further experiments tested various loss functions: MSELoss (baseline), MAELoss, SmoothL1Loss, CrossEntropyLoss, CORN, and CORAL. As summarized in Table 7, regression-based loss functions outperform ordinal and classification-based losses for structure scoring, suggesting that continuous score regression is more suitable for modeling essay structure traits. The comparison across loss functions indicates that regression-based objectives are more suitable for modeling structure traits, as they better preserve score continuity and ordinal relationships. In contrast, ordinal and classification-based losses tend to introduce instability and reduced agreement in this task.

The last experiment compares the Adam and AdamW optimizers, with fixed configurations: learning rate 2e-5 and MSELoss. Table 8 shows that although AdamW achieves a lower final loss, the Adam optimizer yields more stable validation agreement, indicating better generalization for essay structure assessment. Although both optimizers converge effectively during training, the validation results suggest that Adam provides more stable generalization for essay structure assessment compared to AdamW.

Table 7. Comparison of different loss functions for structure scoring, evaluated using QWK on training and validation datasets

Loss function	QWK training	QWK validation	Final loss
MSELoss	0.9997	0.8022	0.0036
CrossEntropyLoss	0.9995	0.7796	0.0371
CORN Loss	0.9972	0.7740	0.2500
MAELoss	0.9376	0.7492	0.3183
CORAL Loss	0.5911	0.4849	3.7874

Table 8. Performance comparison between Adam and AdamW optimizers for structure scoring under fixed hyperparameter settings

Optimizer	QWK training	QWK validation	Final loss
Adam	0.9997	0.8022	0.0036
AdamW	0.9997	0.7949	0.0029

After all experiments were completed, the final evaluation was performed on the testing data using the best configuration: learning rate 2e-5, optimizer Adam, MSELoss, batch size 16, and 50 epochs. The model produced a QWK of 0.7906 on the testing data, a value very close to the validation QWK (0.8022). This indicates that the model has good generalization and does not suffer from overfitting. The validation QWK of 0.8022 represents an improvement of approximately 5.55% compared to the structure model in Tambe and Kulkarni (2022) [1], which reported a QWK of 0.76 on validation data. This improvement percentage was calculated using (5):

$$Improvement(\%) = \frac{QWK_{proposed} - QWK_{baseline}}{QWK_{baseline}} \times 100 \quad (5)$$

where $QWK_{proposed}$ =0.8022 and $QWK_{baseline}$=0.76, yielding:

$$Improvement(\%) = \frac{0.8022 - 0.76}{0.76} \times 100\% \approx 5.55\%$$

The results in Table 9 demonstrate that the proposed structure scoring model maintains consistent performance across training, validation, and testing sets, indicating robustness against overfitting. Representative examples in Table 10 illustrate that the predicted trait-level scores closely align with human annotations, supporting the reliability and interpretability of the proposed trait-based structure scoring approach.

Table 9. Final QWK evaluation results of the structure scoring module across training, validation, and testing datasets

Dataset	Number of essays	QWK
Training	1098	0.9997
Validation	235	0.8022
Testing	236	0.7906

Table 10. Trait-level structure scoring predictions on the test set with human scores and model outputs

Essay id	Essay	Ideas score	Organization score	Style score	Structure true score	Ideas pred score	Organization pred score	Style pred score	Structure pred score
18776	A time when someone else I knew was patient...	3	3	2	11	3	3	2	11
17984	Patience is a virtue. It's something that we all...	2	2	2	8	2	2	2	8
19215	I had a friend who is @CAPS@ best friend a...	3	3	2	11	2	2	2	8
18471	Have you ever have to wait in line to get on th...	2	2	2	8	2	2	2	8

Based on all experiments, the best model used the BERT+BiLSTM+attention pooling architecture with Adam optimizer, MSELoss, and learning rate 2e-5. Although the initial holistic approach resulted in the highest validation QWK (0.8253), the trait-based approach was chosen as the final configuration. This model

not only demonstrates strong performance, but it also supports interpretability and transparency through per-trait scores. This approach is in line with the explainable principle of AES as emphasized by Xue *et al.* [12], which highlights the importance of per-aspect evaluation to support more meaningful learning feedback.

Figure 5 illustrates the distribution of actual and predicted structure scores on the testing data. The similarity in curve shapes indicates that the model successfully captures the dominant scoring patterns, while minor deviations at higher score ranges reflect the inherent difficulty of distinguishing essays with closely ranked quality.

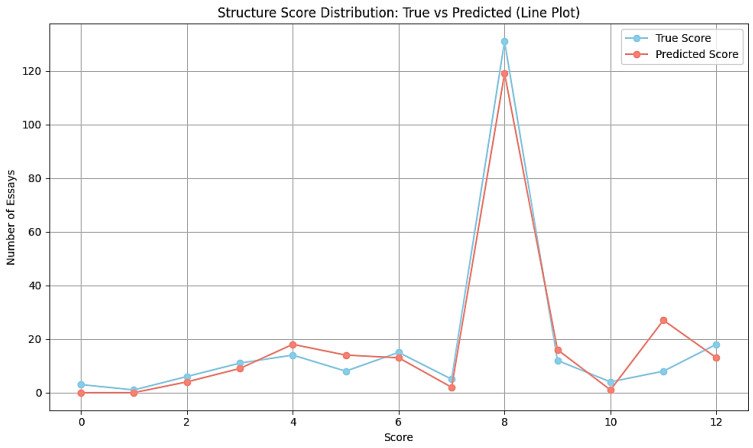

Figure 5. Distribution of actual and predicted structure scores on the testing dataset, illustrating the alignment between model predictions and human reference scores across different score ranges

3.2. Grammar checking model

The grammar checking experiments indicate that moderate learning rates achieve a more balanced trade-off between precision and recall, enabling reliable detection of grammatical acceptability without excessive false positives. The evaluation was conducted using the metrics of accuracy, F1-score, precision, recall, and eval loss, as summarized in Table 11. As summarized in Table 11, a moderate learning rate achieves the most balanced trade-off between precision and recall, enabling effective grammar error detection without excessive false positives.

Table 12 presents the results of using the grammar checker model that was trained on the CoLA dataset and then applied to evaluate essays from the ASAP dataset. The table displays the predicted grammar labels per sentence and clause, indicating the performance of the model in detecting grammatical errors.

Table 11. Evaluation results of the BERT-based grammar checking model on the CoLA dataset using different learning rates

Learning rate	Accuracy	F1-score	Precision	Recall	Eval loss
1e-5	0.8444	0.8941	0.8460	0.9479	0.3615
2e-5	0.8596	0.9013	0.8779	0.9260	0.3322
3e-5	0.8558	0.8976	0.8833	0.9123	0.3275

Table 12. Sentence-level grammar classification results from the CoLA-trained BERT model on ASAP essays

Essay id	Unit type	Unit number	Sentence id	Sentence	Grammar label	Spelling corrections
18471	Sentence	1	1	Have you ever have to wait in line to get on the bigger...	Correct	-
18471	Sentence	2	2	Have you ever wait until all your chores are done befo...	Correct	-
18471	Sentence	3	3	Have you ever had to wait for your little brother or si...	Correct	-
18471	Sentence	4	4	Well, in this story I will tell you about one time that I h..	Correct	-

3.3. Grammar scoring module

Experiments were conducted by evaluating a combination of BERT (CoLA) grammar checker models with varying learning rates (1e-5, 2e-5, and 3e-5) and number of logistic regression iterations (2000 and 3000). The results in Table 13 show that the best configuration is learning rate 2e-5 with 2000 iterations. The results indicate that increasing the number of logistic regression iterations beyond a certain point does not improve performance, suggesting that the grammar scoring model converges efficiently with a moderate iteration setting.

The testing evaluation results are shown in Table 14. The evaluation was conducted using the best configuration, specifically learning rate 2e-5 and 2000 iterations, which previously showed the highest validation performance. Tables 14 and 15 indicate that the grammar scoring model achieves stable agreement across training, validation, and testing sets, demonstrating consistent generalization despite the inherent difficulty of grammar trait prediction.

Table 13. Grammar scoring performance measured using QWK under different grammar checker learning rates and logistic regression iteration settings

Learning rate	Iterations	QWK training	QWK validation
1e-5	2000	0.4055	0.4730
1e-5	3000	0.4055	0.4730
2e-5	2000	0.4031	0.4746
2e-5	3000	0.4031	0.4746
3e-5	2000	0.3537	0.4710
3e-5	3000	0.3537	0.4710

Table 14. Final QWK evaluation results of the grammar scoring module across training, validation, and testing datasets

Dataset	Number of essays	QWK
Training	1098	0.4031
Validation	235	0.4746
Testing	236	0.3923

Table 15. QWK evaluation of grammar scoring after score range normalization (0-3 mapped to 0-10) for comparison with prior studies

Dataset	Number of essays	QWK
Training	1098	0.6166
Validation	235	0.6483

To allow comparison with the approach by Tambe, which used outputs with a modified range of 0-10, the proposed method was also evaluated using the same adjusted range. When the original 0-3 range was mapped to 0-10, the model's performance increased accordingly. The resulting QWK on the validation set was 0.6483, which is very close to Tambe's reported value of 0.65. When evaluated under the same score range as prior studies, the proposed grammar scoring module demonstrates comparable agreement, indicating that its predictive capability remains consistent across different scoring scales. The comparison is summarized in Table 15.

An example visualization of the model prediction results on the testing data is shown in Table 16, which compares the actual and predicted grammar scores for each essay. This table also displays the thirteen linguistic features used as inputs for the logistic regression model.

Table 16. Grammar scoring features and prediction results on the test set with thirteen linguistic features for multinomial logistic regression

Essay_id	STg	SWg	SWg_new	Grammar_Ratio	WTs	Ws	Spelling_Ratio	
18776	21	9	6	0.43	240	1	0	
17984	17	10	10	0.59	147	0	0	
19215	28	17	17	0.61	380	0	0	
18471	14	6	4	0.43	223	0	0	
Essay_id	Words_per_ Sentences	Words_per_ Clauses	CTg	CWg	CWg_new	Grammar _ Ratio	Score	Predicted _ score
18776	11.43	5.45	44	11	15	0.25	3	3
17984	8.65	4.59	32	19	19	0.59	3	2
19215	13.57	6.44	59	29	30	0.49	2	3
18471	15.93	7.19	31	9	10	0.29	3	2

There are 13 linguistic features used in the modeling, which represent the number and ratio of grammar errors and the distribution of text length at the sentence and clause levels. The features include: number of sentences (STg), number of sentences with grammar errors (SWg), sentences still incorrect after correction spelling (SWg_new), ratio of grammar errors per sentence (Grammar_Ratio_Sentence), number of clauses (CTg), clauses with grammar errors (CWg), clauses still incorrect after correction spelling (CWg_new), and ratio of grammar errors per clause (Grammar_Ratio_Clause). In addition, other features include the number of words (WTs), number of misspelled words (Ws), spelling error ratio (Spelling_Ratio), average words per sentence (Words_per_Sentence), and average words per clause (Words_per_Clause).

Hybrid AES approaches integrating neural representations with explicit linguistic features have been shown to improve robustness compared to purely end-to-end neural scoring systems. Linguistic indicators such as syntactic error ratios, sentence complexity, and lexical statistics provide complementary information that helps models better capture grammatical proficiency and improve scoring consistency [32].

Figure 6 illustrates the distribution of grammar scores predicted by the proposed model compared to the reference scores. The concentration around score 2 reflects the dominant grammar proficiency level in the dataset, while reduced predictions at extreme scores indicate conservative behavior in sparsely represented classes.

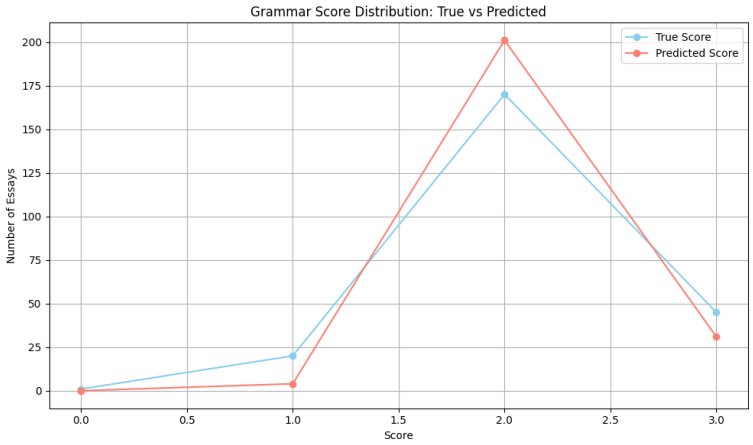

Figure 6. Distribution of actual and predicted grammar scores on the testing dataset, highlighting the model's tendency to concentrate predictions around dominant score categories

The close alignment between the predicted and reference score distributions demonstrates that the proposed grammar scoring model is able to capture the overall grammar proficiency pattern present in the dataset. Although predictions are more concentrated around the dominant score level, slight deviations occur at the lower and higher ends of the score range. This tendency indicates that the model adopts a conservative prediction strategy when handling underrepresented score categories, which helps reduce unstable predictions caused by data imbalance. Despite these minor deviations, the general consistency between both distributions supports the robustness and reliability of the grammar scoring module in assessing essay-level grammatical quality.

3.4. Holistic score

The integration of independently optimized structure and grammar modules results in a stable holistic scoring performance. The combined model maintains substantial agreement across data splits, indicating that modular aggregation does not introduce instability or overfitting.

As shown in Table 17, the integration of structure and grammar modules produces a stable holistic score with substantial agreement, confirming the effectiveness of the proposed modular aggregation scheme. With QWK scores of 0.9760 (train) and 0.8046 (val), the small gap suggests stable performance without overfitting. The validation QWK of 0.8046 also represents an improvement of approximately 3.15% compared to the holistic model in Tambe and Kulkarni's (2022) [1]. research, which reported a QWK of 0.78 on validation data. This improvement percentage was calculated using (5), where $QWK_{proposed} = 0.8046$ and $QWK_{baseline}$=0.78, yielding:

$$Improvement(\%) = \frac{0.8046-0.78}{0.78} \times 100\% \approx 3.15\%$$

Table 18 presents representative examples demonstrating that the aggregated holistic scores closely follow human judgments, further supporting the reliability of the modular AES framework. Figure 7 displays the holistic score distribution on the testing data in the form of a line plot. Both the actual and predicted curves show a peak in the distribution at score 10, indicating that the model is able to maintain the dominant pattern when the two modules are combined.

Table 17. Holistic scoring performance after integrating structure and grammar modules, evaluated using QWK

Dataset	Holistic total QWK	QWK interpretation
Training	0.9760	Almost perfect agreement
Validation	0.8046	Substantial agreement
Testing	0.7847	Substantial agreement

Table 18. Representative comparison of human-assigned and predicted holistic scores on the testing dataset after modular aggregation

Essay_id	Essay	Structure true score	Structure pred score	Grammar true score	Grammar pred score	Total true score	Total pred score
18776	A time when someone else I knew was patient...	11	11	3	3	14	14
17984	Patience is a virtue. It's something that we all...	8	8	3	2	11	10
19215	I had a friend who is @CAPS@ best friend a...	11	8	2	3	13	11
18471	Have you ever have to wait in line to get on th...	8	8	3	2	11	10
18566	@PERSON3 and the @CAPS1 @CAPS2. Vincen...	12	12	3	3	15	15

Figure 7 shows a close alignment between the actual and predicted holistic score distributions after integrating the structure and grammar modules. The consistent peak and similar overall curve shapes indicate stable aggregation behavior, suggesting that the proposed modular framework effectively preserves dominant scoring patterns. Minor deviations are observed at the extreme score ranges, where the model slightly overpredicts scores in the medium range (6-7) and underpredicts the highest score category. These differences are likely influenced by data imbalance at the distribution tails. Despite these deviations, the overall alignment between predicted and reference distributions supports the quantitative results, confirming a substantial level of agreement between the system-generated holistic scores and human judgments.

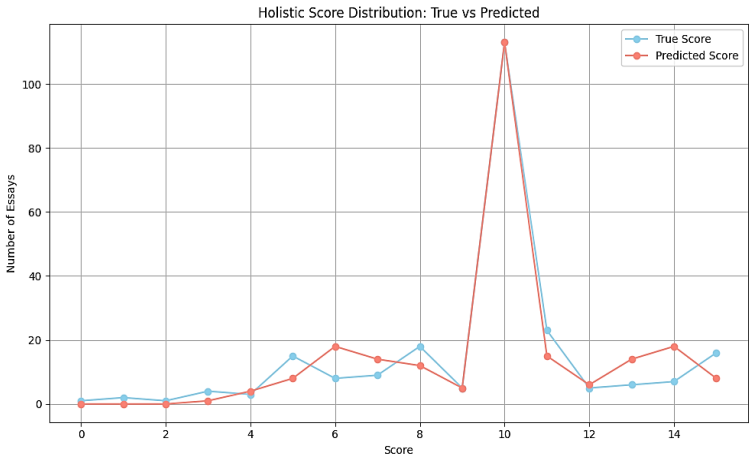

Figure 7. Distribution of actual and predicted holistic scores on the testing dataset, demonstrating the stability of score aggregation after integrating structure and grammar modules

From a practical perspective, the proposed trait-based AES system enables more actionable feedback for educational settings. Instead of providing a single holistic score, the system reports separate scores for ideas, organization, style, and grammar, allowing students to identify specific strengths and

weaknesses in their writing. This supports formative assessment by enabling targeted feedback and revision strategies, particularly in classroom environments where manual trait-level evaluation is time-consuming. Such diagnostic capability aligns with current educational demands for transparent and learner-centered assessment systems.

3.5. Comparison with recent automated essay scoring studies

To contextualize the effectiveness of the proposed modular trait-based AES system, this study compares its performance with recent AES research published between 2019 and 2025. A grammar-aware AES framework proposed in [1] integrates handcrafted grammar features into a neural scoring model and reports a validation QWK of approximately 0.76 for structure-related scoring and 0.65 for grammar scoring. In comparison, the proposed system achieves a higher validation QWK of 0.8022 for structure scoring and a comparable grammar QWK of 0.6483 after score range normalization, indicating stronger agreement with human raters while preserving modular interpretability.

A hierarchical BERT-based transfer learning approach introduced in [12] predicts multiple writing traits using separate subnetworks, achieving strong performance through multi-trait supervision. However, this approach requires fine-grained trait annotations and substantial computational resources to train multiple subnetworks concurrently. Compared to this method, the proposed system adopts a lighter modular design by separating structure and grammar into independent modules, achieving competitive performance with reduced model complexity and improved reproducibility.

Previous studies have also highlighted the transparency limitations of holistic deep learning-based AES models, particularly those based on CNN, LSTM, and end-to-end BERT architectures, which typically produce a single score without providing diagnostic feedback [8]. The trait-based outputs of the proposed system directly address this limitation by generating interpretable scores for ideas, organization, style, and grammar, thereby supporting explainable assessment and more actionable feedback.

Comprehensive surveys on deep learning-based AES indicate that transformer architectures consistently outperform earlier neural models due to their ability to capture long-range dependencies and contextual semantics. However, these models often prioritize predictive accuracy over interpretability, motivating the need for architectures that balance performance and transparency [33].

Recent multi-scale BERT-based AES frameworks demonstrate that jointly modeling sentence-level and essay-level representations can further improve scoring stability across prompts. While such hierarchical architectures often achieve strong predictive performance, they also increase model complexity and training cost. In contrast, the modular design proposed in this study maintains competitive agreement with human raters while preserving architectural simplicity and clearer trait separation [34].

Emerging research also explores zero-shot and rationale-based AES frameworks that aim to generalize across prompts without task-specific retraining. These approaches emphasize adaptability and explanation generation, addressing limitations in prompt dependency and scalability. Incorporating such adaptive scoring mechanisms represents a promising direction for extending modular and trait-specific AES architectures [35].

More recent AES research emphasizes that surface-level and grammar-focused features alone are insufficient to capture semantic and discourse-level quality in student writing [3], [4]. By integrating contextual BERT embeddings with attention pooling in the structure scoring module, the proposed system is able to better capture salient textual information, as reflected in consistent improvements across validation and testing QWK scores.

Overall, compared with recent AES studies, the proposed modular trait-based framework demonstrates a balanced trade-off between performance, interpretability, and computational efficiency. By avoiding heavy hierarchical architectures while maintaining multi-trait assessment, this approach offers a practical and transparent alternative for real-world educational deployment.

4. CONCLUSION

This study confirms that a modular, trait-based AES framework provides a more interpretable and effective alternative to conventional holistic scoring approaches. By separating structure and grammar assessment, the proposed system achieves reliable agreement with human raters while enabling explicit evaluation of individual writing traits, demonstrating stable generalization across validation and testing data. The main contribution of this work lies in the introduction of a practical trait-level scoring framework that balances performance and transparency without requiring complex hierarchical architectures or extensive fine-grained annotations. From an educational perspective, the modular design enhances the diagnostic value of AES by allowing strengths and weaknesses in specific writing aspects to be identified more clearly, supporting more objective and actionable feedback for learners. Nevertheless, the current framework is limited to a predefined set of traits and English-language datasets. Future research may extend this approach

to additional writing dimensions, multilingual AES settings, and tighter integration with classroom learning tools to support formative assessment and adaptive writing instruction.

The proposed structure module achieved a QWK of 0.7906 on testing data, while the integrated holistic score reached 0.7847, indicating stable generalization across data splits. By separating semantic structure evaluation from grammatical assessment, the framework enables clearer diagnostic feedback at the trait level, supporting more transparent and actionable educational assessment. The primary contribution lies in the explicit architectural decomposition of structure and grammar into independently optimized modules, reducing representational interference and enhancing reproducibility. However, grammar scoring performance remains lower than structure scoring, reflecting domain differences between CoLA and learner essay data. Future research should explore learner-specific grammar corpora, joint optimization strategies, and cross-prompt validation to further improve robustness and generalizability in real-world educational settings.

ACKNOWLEDGMENTS

The authors would like to acknowledge the *Lembaga Penelitian dan Pengabdian Masyarakat* (Institute for Research and Community Service), Institut Teknologi Nasional Bandung, for providing financial support for the publication of this article under a research grant/contract scheme.

FUNDING INFORMATION

This research assignment letter is 49b/J.016/LPPM/Itenas/II/2025.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Jasman Pardede	✓	✓		✓	✓	✓	✓	✓	✓	✓	✓	✓		✓
Rizka Milandga Milenio			✓			✓		✓	✓	✓	✓		✓	
Thalita Zharifa Nathania	✓	✓	✓	✓			✓	✓	✓	✓	✓			

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

DATA AVAILABILITY

Data availability is not applicable to this paper as no new data were created or analyzed in this study. The study uses publicly available datasets, namely the ASAP dataset [20] and the CoLA dataset [21], which can be accessed through their respective repositories.

REFERENCES

[1] A. A. Tambe and M. Kulkarni, "Automated essay scoring system with grammar score analysis," in *Proceedings - 2nd International Conference on Smart Technologies, Communication and Robotics (STCR)*, 2022, pp. 1–7, doi: 10.1109/STCR55312.2022.10009053.

[2] N. Zainal and M. H. A. Hassan, "Automated essay scoring (AES) using English essay question," in *2022 IEEE 20th Student Conference on Research and Development (SCOReD)*, 2022, pp. 1–6, doi: 10.1109/SCOReD57082.2022.9973989.

[3] L. B. Das *et al*., "FACToGRADE: Automated essay scoring system," in *Proceedings of the 2022 IEEE International Conference on Industry 4.0, Artificial Intelligence, and Communications Technology (IAICT)*, 2022, pp. 42–48, doi: 10.1109/IAICT55358.2022.9887447.

[4] A. Mizumoto and M. Eguchi, "Exploring the potential of using an AI language model for automated essay scoring," *Research Methods in Applied Linguistics*, vol. 2, no. 2, 2023, doi: 10.1016/j.rmal.2023.100050.
[5] P. Wangkriangkri, C. Viboonlarp, A. T. Rutherford, and E. Chuangsuwanich, "A comparative study of pretrained language models for automated essay scoring with adversarial inputs," in *IEEE Region 10 Annual International Conference, Proceedings/TENCON*, 2020, pp. 875–880, doi: 10.1109/TENCON50793.2020.9293930.
[6] S. Mathias and P. Bhattacharyya, "ASAP++: Enriching the ASAP automated essay grading dataset with essay attribute scores," in *LREC 2018 - 11th International Conference on Language Resources and Evaluation*, 2018, pp. 1169–1173.
[7] S. Prabhu, K. Akhila, and S. Sanriya, "A hybrid approach towards automated essay evaluation based on BERT and feature engineering," in *2022 IEEE 7th International Conference for Convergence in Technology (I2CT)*, 2022, pp. 1–4, doi: 10.1109/I2CT54291.2022.9824999.
[8] F. Li, X. Xi, and Z. Cui, "Automatic essay scoring method based on multi-scale features," *Applied Sciences*, vol. 13, no. 11, 2023, doi: 10.3390/app13116775.
[9] Y. Attali and S. Sinharay, "Automated trait scores for GRE writing tasks," *ETS Research Report Series*, vol. 2015, no. 1, pp. 1–14, 2015, doi: 10.1002/ets2.12062.
[10] E. Mayfield and A. W. Black, "Should you fine-tune BERT for automated essay scoring?," in *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications (BEA)*, 2020, pp. 151–162, doi: 10.18653/v1/2020.bea-1.15.
[11] M. Beseiso and S. Alzahrani, "An empirical analysis of BERT embedding for automated essay scoring," *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 10, pp. 204–210, 2020, doi: 10.14569/IJACSA.2020.0111027.
[12] J. Xue, X. Tang, and L. Zheng, "A hierarchical BERT-based transfer learning approach for multi-dimensional essay scoring," *IEEE Access*, vol. 9, pp. 125403–125415, 2021, doi: 10.1109/ACCESS.2021.3110683.
[13] A. Warstadt, A. Singh, and S. R. Bowman, "Neural network acceptability judgments," *Transactions of the Association for Computational Linguistics*, vol. 7, pp. 625–641, 2019, doi: 10.1162/tacl_a_00290.
[14] H. Do, S. Ryu, and G. G. Lee, "Teach-to-reason with scoring: Self-explainable rationale-driven multi-trait essay scoring," *Expert Systems with Applications*, vol. 319, 2026, doi: 10.1016/j.eswa.2026.132119.
[15] J. Atkinson and D. Palma, "An LLM-based hybrid approach for enhanced automated essay scoring," *Scientific Reports*, vol. 15, 2025, doi: 10.1038/s41598-025-87862-3.
[16] N. A. Kurdhi and A. Saxena, "Evaluating quadratic weighted kappa as the standard performance metric for automated essay scoring," in *Proceedings of the 16th International Conference on Educational Data Mining*, Bengaluru, India, 2023, pp. 103–113, doi: 10.5281/zenodo.8115784.
[17] E. del Gobbo, A. Guarino, B. Cafarelli, and L. Grilli, "GradeAid: A framework for automatic short answers grading in educational contexts—design, implementation and evaluation," *Knowledge and Information Systems*, vol. 65, pp. 4295–4334, 2023, doi: 10.1007/s10115-023-01892-9.
[18] S. V. Chilukoti, L. Shan, V. S. Tida, A. S. Maida, and X. Hei, "A reliable diabetic retinopathy grading via transfer learning and ensemble learning with quadratic weighted kappa metric," *BMC Medical Informatics and Decision Making*, vol. 24, no. 1, pp. 1–12, 2024, doi: 10.1186/s12911-024-02446-x.
[19] L. Chivinge, L. K. Nyandoro, and K. Zvarevashe, "Quadratic weighted kappa score exploration in diabetic retinopathy severity classification using EfficientNet," in *2022 1st Zimbabwe Conference of Information and Communication Technologies (ZCICT)*, Harare, Zimbabwe, 2022, pp. 1–9, doi: 10.1109/ZCICT55726.2022.10045938.
[20] The Hewlett Foundation, "Automated student assessment prize (ASAP): Automated essay scoring," *Kaggle*. Available: https://www.kaggle.com/competitions/asap-aes.
[21] Corpus of Linguistic Acceptability (CoLA), "The Corpus of Linguistic Acceptability." Available: https://nyu-mll.github.io/CoLA/.
[22] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman, "GLUE: A multi-task benchmark and analysis platform for natural language understanding," in *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, Brussels, Belgium, 2018, pp. 353–355, doi: 10.18653/v1/W18-5446.
[23] L. Zhang *et al*., "Sliding-BERT: Striding towards conversational machine comprehension in long context," *Advances in Artificial Intelligence and Machine Learning*, vol. 3, no. 3, pp. 1325–1339, 2023, doi: 10.54364/aaiml.2023.1178.
[24] Y. Nie, "Automated essay scoring with SBERT embeddings and LSTM-attention networks," *PeerJ Computer Science*, 2025, doi: 10.7717/peerj-cs.2634.
[25] S. Lee, Y. Cai, D. Meng, Z. Wang, and Y. Wu, "Unleashing large language models' proficiency in zero-shot essay scoring," in *Findings of the Association for Computational Linguistics: EMNLP 2024*, 2024, pp. 181–198, doi: 10.18653/v1/2024.findings-emnlp.10.
[26] M. Behrendt, S. S. Wagner, and S. Harmeling, "MaxPoolBERT: Enhancing BERT classification via layer- and token-wise aggregation," *arXiv preprint*, 2025, doi: 10.48550/arXiv.2505.15696.
[27] A. Kumar, N. Ware, and S. Gupta, "Leveraging transfer learning: Fine-tuning methodology for enhanced text classification using BERT," in *2024 IEEE Pune Section International Conference (PuneCon)*, 2024, pp. 1–5, doi: 10.1109/PuneCon63413.2024.10895448.
[28] H.-S. Chang, R.-Y. Sun, K. Ricci, and A. McCallum, "Multi-CLS BERT: An efficient alternative to traditional ensembling," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, Toronto, Canada, Jul. 2023, vol. 1, pp. 821–854, doi: 10.18653/v1/2023.acl-long.48.
[29] X. Zou, Y. Hu, Z. Tian, and K. Shen, "Logistic regression model optimization and case analysis," in *Proceedings of IEEE 7th International Conference on Computer Science and Network Technology (ICCSNT)*, 2019, pp. 135–139, doi: 10.1109/ICCSNT47585.2019.8962457.
[30] X. Fan and N. Zhang, "Comparative analysis of sparse multinomial logistic regression and convolutional neural networks for multi-class image classification," in *2025 5th International Conference on Consumer Electronics and Computer Engineering (ICCECE)*, 2025, pp. 44–48, doi: 10.1109/ICCECE65250.2025.10984434.
[31] J. Sun, T. Song, W. Peng, and J. Song, "A survey of automated essay scoring: Challenges, advances, and future," *Neurocomputing*, vol. 650, 2025, doi: 10.1016/j.neucom.2025.130916.
[32] M. Faseeh *et al*., "Hybrid approach to automated essay scoring: Integrating deep learning embeddings with handcrafted linguistic features for improved accuracy," *Mathematics*, vol. 12, no. 21, 2024, doi: 10.3390/math12213416.
[33] H. Misgna, B.-W. On, I. Lee, and G. S. Choi, "A survey on deep learning-based automated essay scoring and feedback generation," *Artificial Intelligence Review*, vol. 58, 2025, doi: 10.1007/s10462-024-11017-5.
[34] 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," in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational*

Linguistics: Human Language Technologies, Seattle, United States, Jul. 2022, pp. 3416–3425, doi: 10.18653/v1/2022.naacl-main.249.
[35] T. Shibata and Y. Miyamura, "LCES: Zero-shot automated essay scoring via pairwise comparisons using large language models," in *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, Suzhou, China, Nov. 2025, pp. 29988–30001, doi: 10.18653/v1/2025.emnlp-main.1523.

BIOGRAPHIES OF AUTHORS

Jasman Pardede earned his bachelor's degree in science and mathematics from Universitas Andalas (Unand), Indonesia, in 2001. He went on to receive his Master of Engineering in Informatics Engineering from Institut Teknologi Bandung (ITB) in 2005, and later completed his doctoral degree in the same field at ITB in 2021. His dissertation was titled "Relevance Feedback by Composite Feedback Objects on CBIR." Since 2005, he has been a lecturer at Institut Teknologi Nasional Bandung, Indonesia. His research interests include image retrieval, data mining, machine learning, and deep learning. He can be contacted at email: jasman@itenas.ac.id.

Rizka Milandga Milenio earned his bachelor's degree in science and mathematics from Universitas Negeri Malang (UM), Indonesia, in 2017. He completed his Master of Engineering in informatics from Institut Teknologi Bandung (ITB) in 2024. Since 2025, he has been a lecturer at Institut Teknologi Nasional Bandung, Indonesia. His research interests include computer vision, smart education, health informatics, machine learning, and deep learning. He can be contacted at email: rizkamilandga@itenas.ac.id.

Thalita Zharifa Nathania is an undergraduate student in the Informatics Study Program at Institut Teknologi Nasional (ITENAS) Bandung, Indonesia. Her academic interests include artificial intelligence, natural language processing, and the development of intelligent systems for educational applications. Her current research focuses on automated essay scoring (AES) in English using BERT-based models to evaluate writing quality based on structure and grammar. This is her first academic publication. In this paper, she contributed to data preprocessing, model architecture design, experimental evaluation, and the interpretation of results. She can be contacted at email: thalitazharifa10@gmail.com.