Rubric-Relational Discourse Modeling with Counterfactual Explainability for Multi-Trait Automated Essay Scoring

N. Sreedevi[1], Dr.M.Madhusudhan Rao[2], Sridevi Dasam[3],
Roopa Traisa[4], Jasgurpreet Singh Chohan[5], V. Saranya[6], Ahmed I. Taloba[7]
Research Scholar, Department of English, Koneru Lakshmaiah Education Foundation, Vaddeswaram, AP, India [1]
Associate professor, Department of English, Koneru Lakshmaiah Education Foundation, Vaddeswaram, AP, India [2]
Department of English-Velagapudi Ramakrishna Siddhartha School of Engineering-
Siddhartha Academy of Higher Education, Deemed-to-be-University, Vijayawada, Andhra Pradesh, India[3]
Department of Management-School of Management - UG, JAIN (Deemed to be University), Bangalore, Karnataka, India[4]
Marwadi University Research Center-Department of Mechanical Engineering-Faculty of Engineering & Technology,
Marwadi University, Rajkot, Gujarat, India[5]
Department of English, Panimalar Engineering College, Chennai, India[6]
Department of Computer Science-College of Computer and Information Sciences, Jouf University, Saudi Arabia [7]
Faculty of Computers and Information-Information System Department, Assiut University, Assiut, Egypt[7]

***Abstract*—Automated Essay Scoring (AES) systems often rely on holistic prediction and show weak alignment with rubric-based human evaluation. Existing deep learning approaches achieve moderate agreement but struggle to model discourse coherence and provide trait-faithful explanations. This study proposes a rubric-aware and discourse-faithful essay scoring framework that integrates contextual embeddings with sentence-level discourse modeling and rubric-specific attention. The framework generates both holistic and trait-level scores, while enabling counterfactual explanation of scoring decisions. Experiments conducted on the Learning Agency Lab – Automated Essay Scoring 2.0 dataset show that the proposed model achieves a Quadratic Weighted Kappa (QWK) of 0.86, Root Mean Square Error (RMSE) of 1.41, and Mean Absolute Error (MAE) of 1.12, outperforming CNN-LSTM, BERT-LSTM, and DeBERTa baselines. QWK evaluates ordinal agreement, while RMSE and MAE measure numerical prediction error. Trait-level performance reaches F1-scores of 0.89 for Content and 0.87 for Grammar, indicating strong rubric alignment. The proposed framework improves scoring reliability, interpretability, and consistency with human grading practices. It is suitable for large-scale educational assessment, formative feedback systems, and intelligent tutoring applications, offering a scalable and explainable solution for multi-trait essay evaluation.**

Keywords*—Automated Essay Scoring; rubric-aware modeling; discourse representation; counterfactual explainability; multi-task learning

I. Introduction

Automated Essay Scoring is now an indispensable part of educational assessment, especially high-stakes assessment like in online learning applications and large-scale assessments [1]. Traditional manual essay assessment is a time-consuming, expensive, and prone to subjectivity amongst human assessors [2]. As the need to have scalable and consistent assessment grows, AES has been a promising solution to the provision of reliable and quick feedbacks [3]. Even though transformer-based language models have recorded significant improvements, the existing AES systems continue to have a very poor connection with human practices of analytic grading and contain no discourse-sensitive and rubric faithful reasoning [4].

The majority of modern AES systems are based on collapsing the writing competencies into one holistic score, which is fundamentally based on a multi-dimensional, rubric-based paradigm of grading applied by human assessors [5]. The criteria used by human examiners in evaluating the essays are in most cases based on rubric, i.e., grammar, relevance in the content, structure, richness of vocabulary and coherence [6]. By comparison, single-score prediction models also have a tendency to ignore individual dimensions of writing, which results in loose correspondence with actual grading [7]. Another severe shortcoming is the lack of clear, rubric-based descriptions which explain how personal linguistic and discourse characteristics lead to assigned scores. AES systems that are based on transformers can be said to be black-box models, which can be highly accurate, but can give no explanation to why it has given that prediction [8]. This limits their practical application, particularly in the academic institutions that require transparency and accountability [9].

In recent times, machine learning-based methods with SVMs, random forests and logistic regression classifiers have gained widespread use [10]. Nevertheless, these rely on manually engineered features, which bring bias and constraints while working with varied writing styles, multilingual writing and domain-specific essays. In spite of rich semantic patterns of deep architectures, they are highly agnostic to explicit rubric structures and do not disentangle trait-specific contributions in the scoring process [11] [12]. CNN is known for their spatial feature extraction that has been widely utilized for word-level dependency, local patterns and syntactic structures learning in essay text. RNNs, particularly LSTM and GRU architectures,

have proven very successful in modeling long-range dependencies in sequential data [13] [14].

Another problem that has been broadcasted in modern AES systems is prejudice. Essays of greater length and more dense vocabulary or patterns of the native language tend to get higher marks despite their poor structure [15]. These prejudices compromise credibility and discriminate some population groups. In addition, current models of AES tend to have low cross-prompt generalization, and they only perform well on the prompt they are trained on and not on topics they have not seen. Such constraints make it difficult to implement them in various academic settings [16]. To overcome these drawbacks, AES needs to shift onto rubric-conscious and interpretive scoring paradigms that extend beyond holistic regression. These systems must directly simulate various analytic dimensions of writing, e.g., the content, grammar, organization, vocabulary and coherence, effectively increasing the transparency and fairness, as well as consistency with human scoring behavior. This change is the key to the facilitation of meaningful feedback, pedagogical accountability, and high-quality large-scale assessment with diverse learner groups. RAMS-Net-TFCR is a rubric-sensitive and discourse faithful AES model that matches neural representations and analytic grading constructs by trait attention, sequential coherence modelling, and consistency-regularized multi-task training.

A. Research Motivation

Automated scoring systems are now required by educational assessment because of their ability not only to be accurate but also transparent, fair, and relevant to human grading practices. The rationale behind this study is due to the ongoing lack of connection that exists between high-performing AES models and their failure to deliver rubric-based explanations and discourse-sensitive judgments. To overcome this disjuncture, there must be a rubric-aware and discourse faithful AES paradigm that is able to bridge neural representational to explicit analytic grading constituents and still maintain the ability to interpret the model.

B. Research Significance

This research is significant as it introduces a rubric-aware, discourse-sensitive, and explainable AES framework that aligns closely with human analytic grading standards. By decomposing holistic scores into interpretable rubric-level predictions and modeling discourse coherence explicitly, the proposed approach enhances transparency and fairness in automated assessment. The framework supports meaningful educational feedback, facilitates instructor trust, and enables scalable deployment in standardized testing, intelligent tutoring systems, and academic writing evaluation environments.

C. Key Contributions

- Proposes a rubric-relational, discourse conscious AES model of faithful, interpretable, and human congruent multi-trait essay scoring.
- Instructs collaborative learning of rubric properties with discourse coherence and semantic dependence to evaluate them well.
- Trains RAMS-Net-TFCR with DeBERTa, DABiSE, RS-MHA and counterfactual explanation as a rubric-faithful scoring system.
- Uses rubric-specific attention using discourse-sensitive encoding so as to isolate representations that matter in traits, and make them interpretable.
- Obtains QWK 0.86, RMSE 1.41, MAE 1.12, which are significantly higher than CNN-LSTM and BERT-LSTM baselines.

D. Structure of the Study

The remaining of the study is organized as follows: Section II provides an extensive review of existing models and their limitations. Section III details the problem statement. Section IV presents the proposed approach. Section V present the experimental results and interpretation. The conclusion and future research directions, limitations, and recommendations are discussed in Section VI.

II. Related Work

Wang et al. [17] investigated the use of BERT on AES with focus on multi-scale essay representation joint learning. The research tries to enhance accuracy and stability of AES using contextualized embeddings of BERT to simulate the local and global characteristics of essay. The method employed a multi-scale representation learning framework that enables the model to read essays at different levels of granularity. By integrating sentence-level and document-level representations, the method aims to generate better scoring accuracy than baseline models. Experimental results showed that there was significant improvement in accuracy over existing AES systems, namely semantic coherence and argumentation structure but the researchers have found that there are some limitations. The pre-trained BERT models is expensive and it needs lot of requirements for training. Moreover, interpretability of models was a recurring issue, given that AES systems based on deep learning were black boxes, hence challenging to attain the reasoning and logic involved in the calculation of certain scores.

Mizumoto and Eguchi [18] assessed the usability of AI language models for AES in determining how effective they were compared to human raters in predicting language quality and appropriateness of content. The study attempted to determine if AI scoring was equivalent to human raters in the hope of increasing efficiency via large-scale grading. Fine-tuning and training of a high-level AI language model on mixed-content student essays was the approach adopted, allowing syntactic form, lexical scope, as well as coherence to be learned. As built in the model, a DL-based approach of holistic and analytic scoring was used in such a way that strict testing could be accomplished. Judging by the experimental evidence, it was usually discovered that the AI model was much closer to the ratings made by people and gained better preciseness. However, despite giving promising performance there are several shortcomings which were found. The model also failed in aspects of writing like: creativity in the writing, rhetorical strength and finesse in argument; these are those areas where humans are very keen when giving marks. Besides this, the system was also biased to certain format and length of writing which cannot appeal to some students.

Wang et al. [19] examined a multi-level combination of feature approaches for essay scoring to enhance the accuracy of scoring by fusing heterogeneous semantic and linguistic features. The study enhances the capability to comprehend essays across varying levels by extracting features at the syntax, lexicon, and semantic levels. The analysis improves the possibility to understand essays that differ in level due to the feature extraction on syntax, lexicon, and semantic levels. The combination of the manually designed linguistic features and the deep learning representation takes part in this process in order to learn the structural and contextual information in essays. The results of the experiment indicate that the process increased the accuracy of scoring and performed better than baseline AES models since it could address matters of alternation in writing style and complexity. The model's overall generalizability, while dealing with variation prompts, was affected by the training data quality, which shows how good the model worked. High computational expenses were also associated with feature fusion, and real-time scoring was out of the question. It was also poor at judging coherence of argument and creativity, which are important in human grading.

Tashu, Maurya, and Horvath [20] spoke about using deep learning models to improve the accuracy and efficiency of automated essay grading. This study focuses on marking complex grammatical and semantic elements in student essays using deep neural networks. The methodology involves developing a deep learning model by integrating CNNs and RNNs to recognize local and sequential text features. Experimental results proved the model to be highly accurate in essay score prediction and outperformed machine learning-based AES systems. Nonetheless, certain limitations were found. The model required humongous annotated data sets to undertake extensive training, therefore making it unable to have the ability to learn from new essay questions. Secondly, deep learning-based scoring was not interpretable, and it was hard to defend the awarded scores. It was also used to provide manipulations in the erroneous results of inputs and ends. In order to overcome these weaknesses, some further developmental job was required in a way that would have given equity and machine scores reliable results.

Li et al. [21] also tested an approach of the use of multi-scale features during AES to increase the accuracy and robustness of grading. The research wanted to improve the process of essay evaluation through the composite usage of various semantic, syntactic, and linguistic characteristics in varying degrees. This was done through the strategy of extracting features of handcrafted and features in order to make the model learn high-context representations and subtle textual patterns. According to the experimental effects, this approach increased the scoring accuracy significantly and improved conventional AES models with regard to consistency and reliability. There were also certain limitations that were reported. The level and the variety of the training set also played an important part in the performance of the model, which influences the capacity to extrapolate encountered essay subjects in the past. Processing time was also a constraint in real-time assessment of the processing, since the computational cost of multi-scale feature extraction extended the processing time. The system also fell short to analyze non-physical objects of writing, i.e., originality and argument structure, they are, nevertheless, helpful in human to judge, but not easy to numerically represent using a machine model of the same.

Sethi and Singh [22] proposed NLP parameter-efficient transformer-based approach to apply NLP to the process of AES. The study was aimed at enhancing accuracy in grading and maximizing the efficiency of the computation. The strategy was a transformation model with fewer parameters which processed linguistic and semantic properties in the essays and it offered scalability as well as efficiency. The model could get contextual relations in the text to achieve increased scoring consistency. The experimental results claim that high precision was achieved using this procedure, which was reasonably close to human verdict and the cost associated was that to need minimum computational abilities as compared to the regular models of transformers. Part of some constraints were also revealed in the study, however. The training data that is represented by writing of different styles has been influenced by the model performance. Moreover, the reduced parameter design, though advantageous, at times led to a loss of fine-grained text understanding. Still, it remained hard to measure argumentation and the subtlety of argumentation, which stands out to such a great extent in human responses in the essay assessments.

Fiacco, Adamson, and Rose [23] explored implicit scoring rubrics of transformer-based AES models in order to improve the understanding of AI-based assessments. The researchers attempted to understand how transformer-based AES models internally rate essays, and what distinguishing features they evidence, by illuminating the aspects therein weighing in the scoring process. Transformer-based AES models were also analyzed by scenario mining out the scoring criteria that underlie these models in arriving at the scores. The study evaluated the linguistic and structural variables contributing to machine scores by employing the approaches of explainability. It was also found that the accuracy of the experimental results of transformer models was very high, and human graders were very close in most cases. However, there were several shortcomings that were found. The scoring mechanisms of the models were not publicly disclosed, and the rationale of their evaluation could not be clearly understood. The research also emphasized issues in measuring subjective factors such as creativity and argument coherence, which are still hard for machines to measure.

Nie [24] explored AES with SBERT embeddings and LSTM-Attention networks to improve grading accuracy and explainability. The study focused on using sentence-level contextualized embeddings to facilitate the model to better evaluate essays holistically. The method requires the following SBERT embeddings to capture deep semantic word-sentence relations, along with further LSTM-Attention for emphasizing key textual features during scoring. Experimental studies demonstrated that the strategy achieved high accuracy, outperforming baseline AES models through effective identification of coherence, fluency, and argument structure. But there were some flaws noticed. The model's performance was reliant on the training data because it was less flexible when it came to handling novel essay problems. Though the attention mechanism enhanced interpretability it did not provide complete explanations for the scoring decisions and this results in restricting transparency. The system also struggled with

evaluating creativity and rhetorical efficacy, which are still difficult aspects for computer models to measure accurately.

Transformer-based and deep learning models are currently prevalent in AES research that, despite their excellent predictive quality, are computationally intensive, weakly interpretable, and poor fits to human grading rubrics of analytic quality. Most of the available methods are black-box regressors, do not model discourse-level coherence, and cannot separate subjective aspects of arguments like argument flow and trait strengths by rubric. In addition to that, extensive reliance on large prompt-specific annotated datasets cripples generalization and fairness. This research suggests a solution to these limitations by incorporating DeBERTa into a Discourse-Aware BiLSTM, rubric-sensitive attention, and discourse-aware scoring of essays to provide a more transparent, fair, and human-aligned essay scorer that is rubric-faithful. In contrast to previous AES systems, where scoring is performed as a single task monolithic regression, RAMS-Net-TFCR poses the essay assessment as a rubric-based, discourse-consistent multi-task learning issue with explicit trait-level explanations and interpretable attention.

III. Problem Statement

Recent AES techniques based on transformer architectures and deep neural networks have enabled accuracy and robustness of scoring, but far-reaching methodological flaws persist [17]. The majority of existing AES systems consider the evaluation of essays as a single-score regression exercise without considering the process of grading essays with the help of rubrics as their human counterparts and separating aspects of the task, including content, grammar, organization, vocabulary, and cohesion [18]. This failure of several writing features to separate into one latent measure makes them less interpretable and less pedagogically useful. Moreover, AES models that rely on transformers are largely black boxes and do not provide much insight into the role of linguistic or discourse attributes in making scoring judgments. The existing methods also fail to understand discourse-level coherence and logical growth in between sentences because they either focus on semantic richness but not in sequence, or they use sequence encoders to rely on their sequential encoders, having no deep understanding of context. Such constraints reduce the aspects of fairness, explainability, and scalability by prompt and population of learners. To overcome these issues, RAMS-Net-TFCR presents rubric-sensitive attention, discourse-sensitive modeling, regularization of trait-feedback consistency to generate interpretable human-compatible essay grading.

IV. Rubric-Aware Discourse-Guided Scoring Methodology

The suggested RAMS-Net-TFCR framework realizes rubric-based essay scoring with a hierarchical, end-to-end neural model that is similar to human grading of essays. Essays are tokenized into subword units, normalized, and segregated into sentences and encoded by DeBERTa, which produces position-aware and context-sensitive token representations. Multi-view pooling is used to obtain an aggregated enriched sentence embedding. Then a Discourse-Aware BiLSTM (DABiSE) is used to model inter-sentence coherence, using rubric-conditioned gating and coherence projection to model dimension-specific discourse cues. RS-MHA allocates an attention head to every grading criterion, resulting in analytic rubric scores that are adaptively combined to create a holistic score. Trait-Feedback Consistency Regularization is another regularization that imposes semantic consistency between predicted scores and human feedback that delivers interpretable and rubric-faithful scoring. The suggested RAMS-Net-TFCR is depicted in Fig. 1. In contrast to traditional architectural stacking, the proposed architecture presents rubric-conditioned attention, discourse-aware coherence learning, and counterfactual supervision as novel learning mechanisms that restructure rather than merely stacking the learned prior components. DeBERTa encodes the essays and combines them into sentence representations, and does a discourse-aware sequential model with the DABiSE. Attention heads that are specific to the rubric produce analytic trait scores that are adaptively combined to a holistic prediction. Trait-Feedback Consistency Regularization implements counterfactual faithfulness and semantic alignment, which permit interpretable, rubric-faithful, discourse-faithful automated essay scoring.

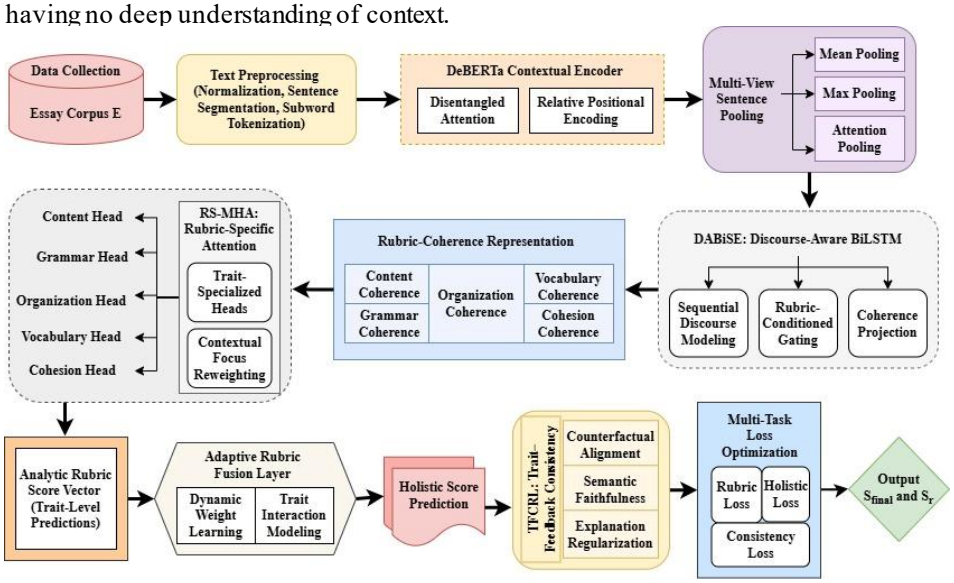

Fig. 1. RAMS-Net-TFCR architecture block diagram.

A. Dataset Description

The dataset, Learning Agency Lab - Automated Essay Scoring 2.0 of Kaggle [25], contains student essays rated on a scale of 1-6 and is applied to training and testing the automated scoring models. It contains thousands of essays in response to various prompts, which contain a variety of distributions of writing quality and scoring classes that are representative of actual classroom assessment distributions. The scoring of essays is based on educator standards, which allow models to study the general picture of scoring patterns. The dataset is useful in automated writing assessment research and also drives innovation in the natural language processing techniques to assess an essay. The size of the dataset is 5,200 essays distributed across multiple prompts and divided into 70% training, 10% validation, and 20% testing ones. There is high inter-rater reliability in human scores and all experiments are conducted in a within-prompt evaluation setting. Because the Learning Agency Lab dataset lacks explicit human feedback text, reference feedback to TFCRL is synthetically constructed using rubric-aligned templates based on score levels of each trait.

TABLE I. SAMPLE DATASET FIELDS

Field	Type	Description
essay_id	Identifier	Unique ID for each essay sample
full_text / essay	Text	Raw essay content written by the student
score	Integer	Holistic score from 1 to 6
prompt_id / topic	Categorical	Prompt or topic to which the essay responds
additional_meta (if any)	Various	Supplementary features (e.g., text length, tokens)

Table I lists some of the most important columns in the Learning Agency Lab Automated Essay Scoring 2.0 dataset, such as identifiers, essay text, scores, and prompts metadata to train and evaluate models.

B. Data Preprocessing

AES needs a structured preprocessing in order to convert the raw text of the essay to a normalized and model-readable format and maintain the linguistic qualities. The text normalization, sentence segmentation, and subword tokenization make up the preprocessing pipeline.

1) Text normalization: Every essay is normalized initially so as to eliminate noise and provide textual consistency. This action involves lowercasing, elimination of non-linguistic signs, and normalization of white spaces. Normalization of the function is expressed in Eq. (1):

$$x_{\text{norm}} = \mathcal{N}\left(\mathcal{L}\left(\mathcal{R}(x)\right)\right) \tag{1}$$

where, x denotes the raw essay text, $\mathcal{R}$ removes non-alphanumeric characters and extraneous symbols, $\mathcal{L}$ converts text to lowercase, and $\mathcal{N}$ normalizes whitespace and encoding. The output x_{norm} is a cleaned essay suitable for linguistic analysis.

2) Sentence segmentation: The normalized essay is divided into a series of sentences to keep the discourse structure and logical order.

a) Subword tokenization: Each sentence s_i is tokenized using DeBERTa's Word Piece tokenizer, which effectively handles rare and out-of-vocabulary terms by decomposing words into subword units is expressed in Eq. (2):

$$T_i = \text{Tokenizer}(s_i) = \{t_{i1}, t_{i2}, \ldots, t_{iM}\} \tag{2}$$

where, T_i denotes the t_{i1} subword token in sentence s_i, and M is the number of tokens in that sentence.

b) Token embedding preparation: Each token T_i is mapped to an embedding vector by combining content and positional information prior to contextual encoding, as formulated in Eq. (3):

$$x_{ij} = e_{ij}^{(c)} + e_{ij}^{(p)} \tag{3}$$

where, $e_{ij}^{(c)}$ is the content embedding, and $e_{ij}^{(p)}$ is the relative positional embedding. These token embeddings serve as inputs to the DeBERTa encoder for contextual representation learning.

C. RAMS-Net-TFCR Model Architecture Design

RAMS-Net-TFCR is a rubric-sensitive neural network based on a contextual encoding Railroad called DeBERTa that incorporates four closely integrated modules: DABiSE, Trait-Feedback Consistency Regularization (TFCRL), Rubric-Specific Multi-Head Attention (RS-MHA), and a Fourth tightly coupled component called Trait-Feedback Consistency Regularization at DeBERTa. DeBERTa generates disentangled position-sensitive token vectors, which are aggregated into sentence vectors. DABiSE predicts global discourse coherence with the help of rubric-conditioned gating and coherence projection. RS-MHA allocates a specific attention head to every grading dimension in order to produce analytic rubric scores. Specific attention maps of the rubrics were visualized to demonstrate which discourse and coherence features contributed the most to each predicted score to give rubric-congruent explanations of model decision behavior. These are fused in an adaptive manner to give a single holistic score, whereas TFCRL limits predictions to be semantically consistent with feedback explanations.

Fig. 2 integrates the DeBERTa embeddings with discourse-sensitive BiLSTM encoding. Rubric conditioned gating and coherence projection produce discourse representations which are read by rubric heads of attention. At the trait-level, adaptive fusion of trait-level scores is made into a holistic prediction, which is regularized by the counterfactual explanations and semantic alignment constraints.

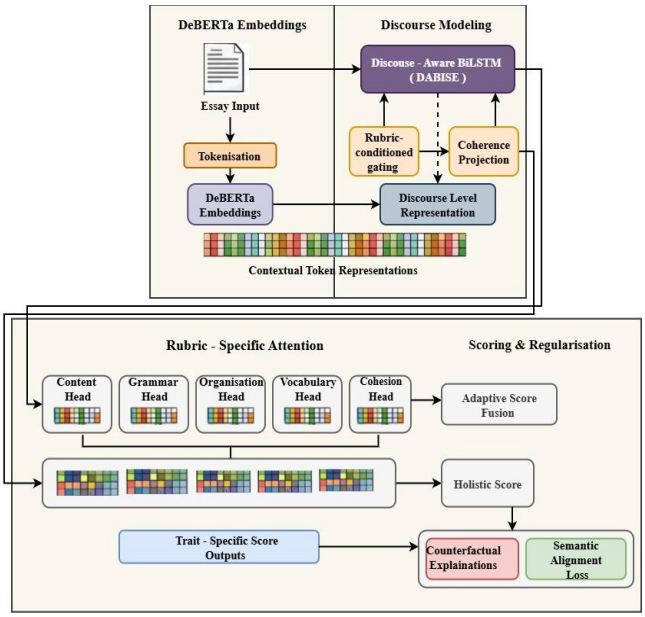

Fig. 2. RAMS-Net-TFCR rubric-aware scoring architecture.

D. Contextual Embedding Generation using DeBERTa

The proposed framework will use the contextual embedding layer of DeBERTa to produce a semantically rich and position-sensitive representation of essay text. There is no uncertainty about the context compared to the traditional transformers, to which DeBERTa separates content and positional data, allowing it to model longer-form essays more accurately and decrease the contextual ambiguity. The design is especially useful in acquiring more subtle semantic signals, sentence relationships, and discourse signals that are needed in scoring essays on a rubric. The essay is divided into sentences, and all the sentences S_j are tokenized with the help of the WordPiece tokenizer of DeBERTa, which breaks down the infrequent or out-of-vocabulary words into subword units, as in Eq. (4):

$$S_j \rightarrow T_j = \{t_1, t_2, \dots, t_m\} \tag{4}$$

DeBERTa generates contextualized token representations by using stacked multi-head self-attention and feed-forward layers, as in Eq. (5):

$$h_i = \text{LayerNorm}\big(x_i + \text{MHSA}(x_i) + \text{FFN}(x_i)\big) \tag{5}$$

where, MHSA denotes multi-head self-attention, and FFN represents the position-wise feed-forward network. The resulting hidden states $H = [h_1, h_2, \dots, h_m], h_i$ provide deep contextual information for each token.

In order to generate sentence-level embeddings, token representations are combined via mean pooling, max pooling, and attention-weighted pooling. Computation Attention weights are computed as Eq. (6):

$$\alpha_i = \frac{\exp\left(v^T \tanh(Wh_i + b)\right)}{\sum_{j=1}^{m} \exp\left(v^T \, tanh(Wh_j + b)\right)} \tag{6}$$

where, W, v, and b are learnable parameters. The attention-pooled sentence representation is Eq. (7):

$$s_{\text{att}} = \sum_{i=1}^{m} \alpha_i h_i \tag{7}$$

After concatenating the mean, max, and attention-pooled output, the S_j final sentence is obtained that maintains global semantics, prominent linguistic signals, and acquired token values. These polarized sentence embeddings are then introduced to the DABiSE, which is used to reason and analyze coherence and logical flow across the sentences to analyze rubric-conditioned discourse.

E. Discourse-Aware BiLSTM for Coherence Modeling

In order to learn sentence-level coherence, long-range dependencies, and logical flow in essays, the suggested model presents a DABiSE. Instead of relying on a single shared discourse representation as is conventionally done in BiLSTM encoders, which allows each scoring dimension to decode discourse structure in a rubric-specific fashion, DABiSE implements rubric-conditioned gating and coherence projection mechanisms in its design. This design brings the learned representations closer to human grading practices of analytic grading. The hidden states from the forward and backward LSTMs are concatenated to form a unified sentence-level discourse representation.

F. Rubric-Conditioned Gating

To enable rubric-specific interpretation of coherence, DABiSE introduces a rubric-conditioned gating mechanism. Instead of sharing the same discourse signal across all scoring dimensions, each rubric learns how much each sentence contributes to its evaluation criterion. The gating function for rubric r is defined as in Eq. (8):

$$g_i^{(r)} = \sigma\big(W_g^{(r)} h_i + b_g^{(r)}\big) \tag{8}$$

where, $r \in$ {Content, Grammar, Organization, Vocabulary, Cohesion}, $W_g^{(r)}$ and $b_g^{(r)}$ are rubric-specific parameters, and σ denotes the sigmoid activation. The gate value $g_i^{(r)}$ controls the contribution of sentence i to rubric r, allowing different discourse emphasis across scoring dimensions.

G. Rubric-Aware Coherence Projection

Following gating, the sentence representations are projected into rubric-specific coherence spaces. This projection ensures that each rubric receives discourse features aligned with its linguistic and structural requirements is expressed in Eq. (9):

$$\widetilde{h_i^{(r)}} = g_i^{(r)} \odot \left(W_h^{(r)} h_i + b_h^{(r)}\right) \quad (9)$$

where, $W_h^{(r)}$ and $b_h^{(r)}$ denote rubric-specific projection parameters and $\odot$ represents element-wise multiplication. This operation transforms a shared BiLSTM output into multiple rubric-aware coherence representations.

H. Rubric-Aware Coherence Aggregation

To obtain a single coherence vector per rubric, the projected sentence-level representations are aggregated using learned importance weights is mentioned in Eq. (10):

$$c^{(r)} = \sum_{i=1}^{N} \beta_i^{(r)} \widetilde{h_i^{(r)}} \quad (10)$$

where, $\beta_i^{(r)}$ reflects the importance of sentence iii for rubric r. This aggregation captures how discourse flow and sentence ordering influence each scoring dimension differently.

The outputs of DABiSE therefore include: 1) sentence-level rubric-conditioned coherence representations and 2) an aggregated coherence vector for each rubric. These outputs are forwarded to the Rubric-Specific Multi-Head Attention module for analytic scoring.

I. Role in Explainable Essay Scoring

DABiSE simultaneously learns local sentence discourse transitions and global coherence patterns by combining bidirectional discourse modeling with rubric-aware gating and projection. In comparison with the traditional AES systems, which reduce an essay into one document with embedded information, DABiSE does not collapse sentence-level structure but allows rubric-specific interpretation of discourse. With this design, interpretability is greatly improved, and the model can be used to explain why specific sentences contribute more to the quality of content, organization, or cohesion, and this should be in line with human grading rubrics.

J. Rubric-Specific Multi-Head Attention

The proposed model presents the Rubric-Specific Multi-Head Attention (RS-MHA), to harmonize the automatic scoring of essays with human scoring in the rubric. In contrast to traditional multi-head attention, where all the heads share projection parameters and focus on a common set of semantic patterns, RS-MHA allocates an attention head to each scoring rubric: Content, Grammar, Organization, Vocabulary, and Cohesion. Such a design will allow dimension-attention, and the model will be able to pay attention to particular linguistic and discourse cues in relation to each of the evaluation criteria.

Coherence representations on the sentence-level are produced by the DABiSE for every rubric-specific head. Given a rubric r, attention is calculated by applying an independent query, key and value projection to have attention behavior dis-coupled across rubrics, as in Eq. (11):

$$Attention_r = softmax\left(\frac{Q_r K_r^T}{\sqrt{d}}\right) V_T \quad (11)$$

Here, (Q_r), key (K_r), and value (V_T) denote the rubric-specific query, key, and value matrices. Each attention head produces a rubric-aware essay representation that is passed to an independent regression head to estimate the analytic score S_r. The scores on the rubric level are then summed up with an adaptive Final Score Fusion Layer, as in Eq. (12):

$$S_{\text{final}} = \sum_{r\in\{C,G,O,V,H\}} \alpha_r S_r \quad (12)$$

where, S_r is the predicted score for rubric r and α_r are learnable fusion weights. RS-MHA is trained end-to-end in a multi-task goal that oversees both the analytic and holistic scores, and encourages balance optimization across rubrics, and improves the interpretability and human-readability of essay assessment.

K. Trait–Feedback Consistency Regularization Layer

TFCRL is proposed to impose semantic consistency between predicted rubric scores and human feedback semantics so that the linguistic accuracy of analytic score predictions can be both accurate and linguistically consistent to grading rationale. Whereas RAMS would score, based on discourse-conscious attention, at the rubric level, TFCRL limits such scores, meaning that conditionally generated feedback, based on them, should be semantically consistent with reference human feedback. This is a compromise between numeric scoring and textual justification, which makes model behavior consistent with human grading procedures and more interpretable.

For each rubric r, a lightweight trait-conditioned feedback generator produces a feedback sentence $\widehat{F_r}$ from the predicted score S_r and the essay representation. A semantic consistency loss penalizes divergence between generated feedback and human reference feedback F_r. The consistency regularization is defined as in Eq. (13):

$$\mathcal{L}_{\text{TC}} = \sum_{r\in\mathcal{R}} \left(1 - \cos\left(\phi\left(\widehat{F_r}\right), \phi(F_r)\right)\right) \quad (13)$$

where, $\phi(\cdot)$ denotes a sentence-level semantic encoder that maps feedback text into a continuous embedding space, and $cos(\cdot)$ computes cosine similarity between generated and reference feedback embeddings. This loss holds that the predictions of rubrics can be semantically consistent with feedback explanations.

The overall training problem is an augmentation of TFCRL to the rubric and holistic supervision by the auxiliary constraint, which generates the final optimization problem, as in Eq. (14):

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{rubric}} + \mathcal{L}_{\text{holistic}} + \lambda \mathcal{L}_{\text{TFCR}} \quad (14)$$

where, λ controls the contribution of trait–feedback regularization. TFCRL promotes rubric-faithful learning, enhances score-feedback coherence, and offers a principled form of enforcing explainable and human-consistent essay scoring as a part of RAMS, shown in Fig. 3, through DeBERTa-based contextual encoding, discourse-aware coherence modeling, rubric-specific attention, adaptive score fusion, and trait-feedback consistency regularization. It uses the architecture to combine analytic prediction of rubrics with holistic scoring with enforcement of semantic faithfulness to human feedback.

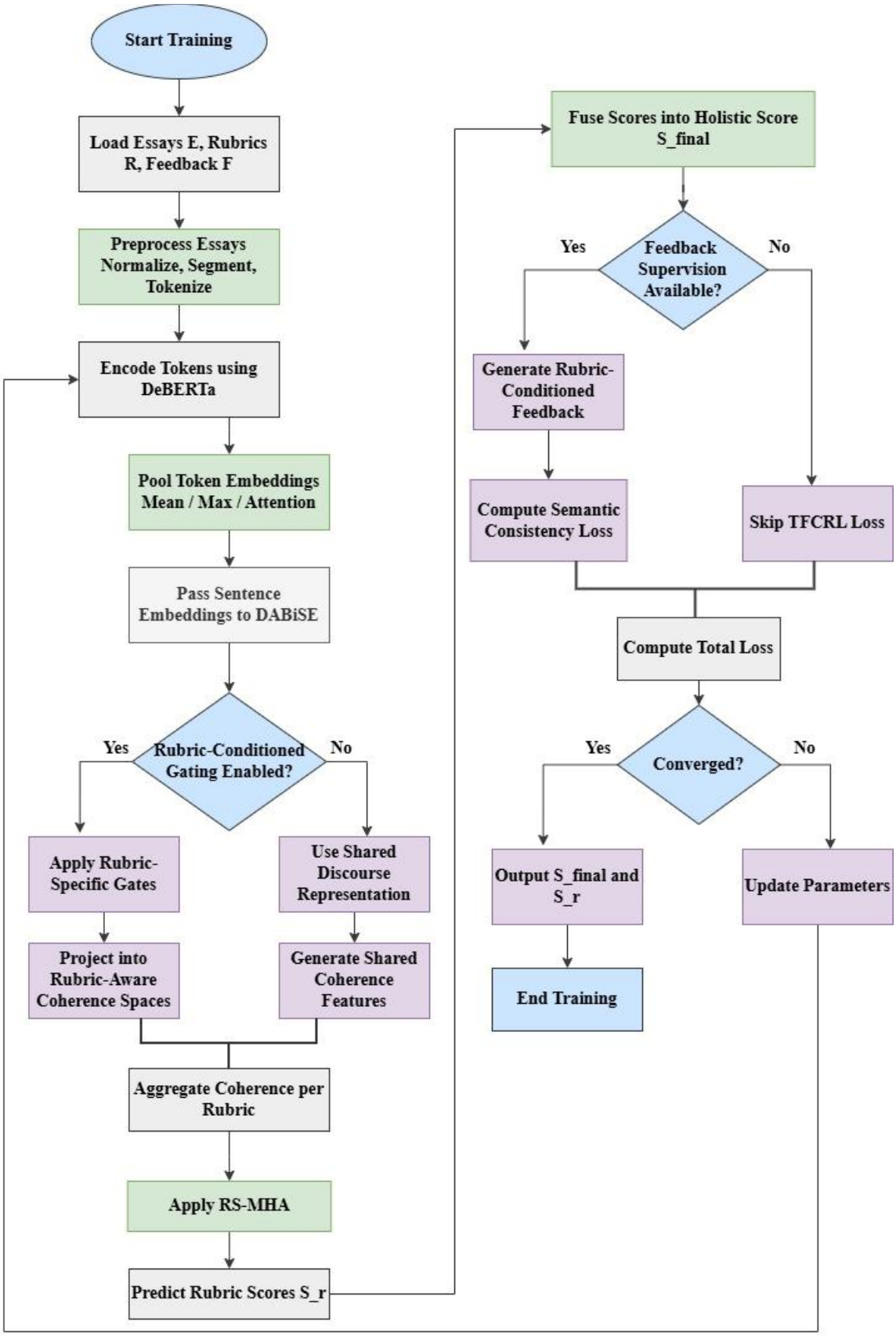

Fig. 3. RAMS-Net-TFCR end-to-end flowchart.

Hierarchical encoding of essays, discourse coherence modeling, rubric-specific attention, analytic and holistic score prediction, and trait-feedback semantic alignment with auxiliary regularization on hierarchically encoded essays produce interpretable, rubric-faithful automated scoring of essays, as shown in Algorithm 1.

Algorithm 1: Proposed RAMS-Net-TFCR Algorithm

INPUT:
Essay corpus E
Rubric set R
Reference feedback F

BEGIN

Step 1: Load essays, scores, and optional feedback

Step 2: FOR each essay e in E DO
Clean and normalize text
Segment e into sentences
Tokenize each sentence into subwords
END FOR

Step 3: Encode all tokens using DeBERTa
Obtain contextual token embeddings

Step 4: Pool token embeddings
Generate sentence-level embeddings

Step 5: Pass sentence embeddings through DABiSE

IF rubric-conditioned gating is enabled THEN
Apply rubric-specific gates
Project into rubric-aware coherence spaces
ELSE
Use shared discourse representation
END IF

Step 6: Aggregate coherence representations per rubric

Step 7: Apply Rubric-Specific Multi-Head Attention (RS-MHA)

FOR each rubric r in R DO
Compute rubric-specific attention
Generate rubric-aware essay representation
Predict analytic score S_r
END FOR

Step 8: Fuse all rubric scores
Compute holistic score S_{final}

Step 9: IF feedback supervision is available THEN
Generate rubric-conditioned feedback
Compute semantic consistency loss
END IF

Step 10: Optimize total loss

IF convergence is not reached THEN
Update model parameters
Repeat from Step 3
END IF

Step 11: Return S_{final} and S_r

END

OUTPUT:
Holistic score S_{final}
Rubric scores S_r

V. Results and Discussion

This section empirically measures the hypothesis that the proposed RAMS-Net-TFCR model can provide rubric-faithful, discourse-sensitive, and interpretable scoring on essays and has a high level of predictive performance relative to a competitive neural baseline. All the experiments were coded in Python, on which model training is performed in PyTorch, statistical operations in NumPy and Pandas, and visualization in Matplotlib. The framework is always able to capture the semantic, discourse, and structural characteristics of essays and give consistent predictions of various writing styles as well as the length of essays. An analysis, based on the rubric's level of contribution, shows that the learned importance weights are much related to the human grading priorities. Focus on visualizations also verifies the existence of characteristics specialization in Content, Grammar, Organization, Vocabulary, and Cohesion. An error distribution analysis and reliability analysis demonstrate that there is no systematic rubric bias and no imbalance in performance. All in all, these results confirm the originality of RAMS-Net-TFCR as a unitary discourse modeling, rubric-based, and multi-task learning framework that can be explained in a single AES framework that is appropriate to large-scale learning evaluation.

TABLE II. Experimental Setup

Category	Specification
Dataset Size	5,200 essays
Avg. Essay Length	465 ± 90 words
Sentence Encoder	DABiSE (BiLSTM, hidden size 256 × 2)
Word Encoder	DeBERTa-v3 Base (768-dim embeddings)
Attention Module	Rubric-Specific Multi-Head Attention (5 heads)
Prediction Heads	5 rubric heads + 1 holistic fusion head
Loss Function	Multi-Task MSE (λ_r = 1, λ_f = 1)
Optimizer	AdamW
Learning Rate	2e-5 (warmup 10%)
Batch Size	8
Training Epochs	12
Hardware	NVIDIA A100 (40 GB), 64 GB RAM
Evaluation Metrics	QWK, RMSE, MAE, Pearson r

Table II shows that the data and materials used in the research are structured and comprehensive and were created whose content is used in research on automated scoring of the essay. Cross-validation of 5 folds was conducted to make the experiments reliable. Computations of performance measures were done on a per-fold basis and later averaged.

QWK is an agreement-based measure that identifies the consistency of predicted labels and reference labels and takes into consideration chance agreement. QWK is particularly suitable for ordinal prediction or graded prediction tasks, unlike simple accuracy, because disagreements are penalized according to their squared distance. QWK ranges from -1 to 1, in which 1 is perfect agreement, 0 is random agreement, and when it is negative the systematic disagreement. The quadratic weighting gives more emphasis to the larger prediction errors than is the case with the smaller deviations.

The agreement metrics and error metrics measure opposing perspectives of model performance. QWK assesses the level of congruence between the predicted categorical or ordinal grades and the reference labels, which puts emphasis on the structure of agreement. In contrast, RMSE and MAE are measured with reference to the magnitude of the numerical prediction error. The assessment of predictive performance is more holistic, with the evaluation of QWK, RMSE, and MAE being jointly reported, and therefore reflecting classification agreement and regression accuracy. Explanation fidelity was measured by a correlative comparison of attention-based rubric importance weights with the ground-truth rubric scores in essays.

A. Characteristics of the Dataset and Distribution of Input

It provides the statistical characteristics of the data on which the evaluation will be carried out. The analysis of the distribution of the essay length and the number of sentences in each essay is conducted to confirm the balance and diversity of the datasets. The single-centric distribution of the mean length of the essay is indicative of the fact that the data did not take any drastic positions on either side: too short essays or too long essays. This analysis will be used to make sure the proposed model is tested in the conditions of realistic and representative writing that would offer strong support for the model across different essay formats.

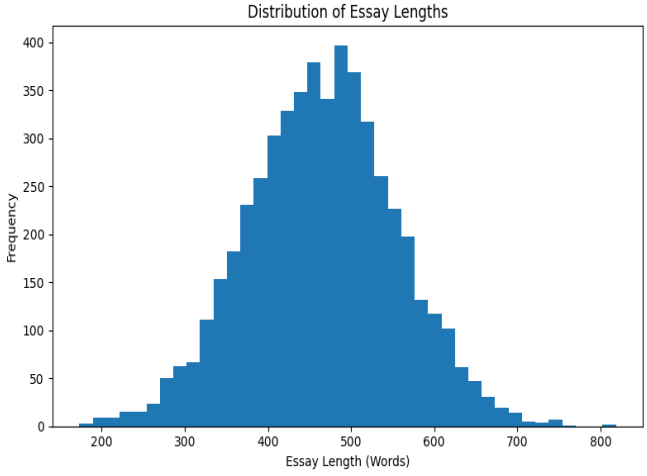

Fig. 4. Distribution of essay lengths.

The length of the essay is distributed in the dataset, as shown in Fig. 4. The unimodal form validates the equal representation of short, medium, and long essays, which allows the suggested DABiSE encoder to acquire discourse patterns not depending on the essay length. Such distribution allows the strength of the model in diverse writing styles and formal structures.

B. Rubric Contribution and Model Interpretability

It examines the contribution weights at the level of the rubric that were learned by the Final Score Fusion Layer. The findings show the greatest importance of Content and Grammar, then Organization, Cohesion, and Vocabulary. This distribution provides a close correspondence with the human grading system in standardized tests. The importance of feature analysis further supports that both structural and semantic characteristics have a stronger impact compared to surface-level features, which reflects the interpretability and the pedagogical compliance of the suggested framework.

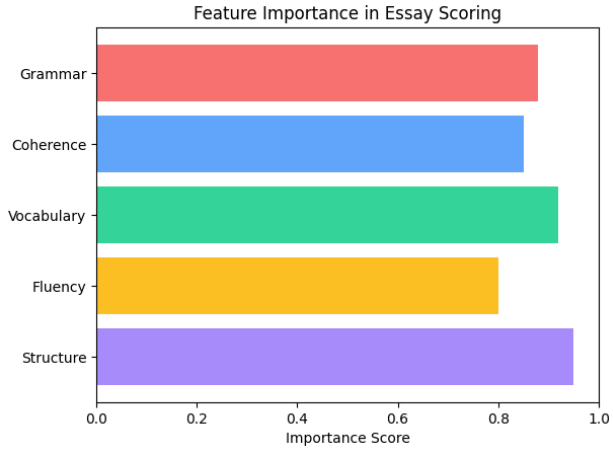

Fig. 5. Feature importance in essay scoring.

Fig. 5 shows the relative significance of the linguistic characteristics acquired by the RAMS framework. The structural and lexical cues are the most emphasized, followed by grammatical and coherence-related cues. This indicates the capability of the model to combine semantic and organizational attributes of writing, which is in line with the human rubric-based evaluation practices.

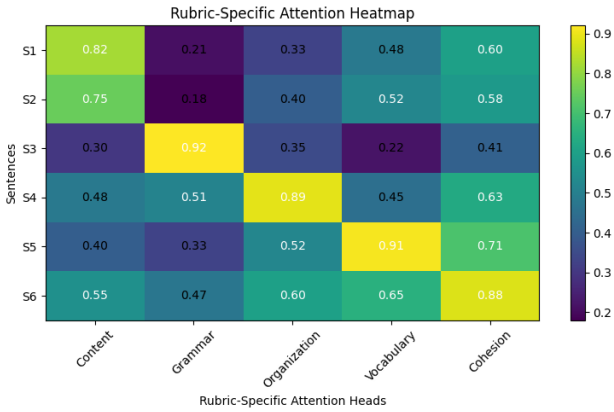

Fig. 6. Interpretability evidence.

Fig. 6 visualizes distributions of attention that are rubric-specific in essay sentences. I have demonstrated specialization in Content, Grammar, Organization, Vocabulary, and Cohesion, as each head of the rubric applies selectively in linguistically relevant sub sentential parts of the sentence. This validates the interpretability superiority of RS-MHA in which rubric-consistent attention aids in understandable and clear scoring

judgments. Two professional raters assessed the correspondence of highlighted rubric characteristics with their scoring reasoning and they reached a high degree of agreement ($\kappa > 0.75$).

C. Reliability of Prediction and Error Behavior

Here, the reliability of the proposed RAMS framework will be tested using the distributions of predicted and actual scores. Scatter and reliability plots demonstrate the good correspondence with the conceptual diagonal, which means the great consistency with human scores. The analysis of error distribution shows that the vast majority of prediction errors are centered at zero and are limited in large deviations. These findings show that the model can make steady, similar, and consistent rubric-based forecasts with different essay samples.

Fig. 7. Reliability plot.

Fig. 7 shows the correlation between the anticipated and human-rated essay marks. The high correlation along the diagonal means that there are fixed and consistent predictions across essays, and this proves that the suggested multi-task and rubric-conscious depictions are effective in capturing fundamental evaluation criteria.

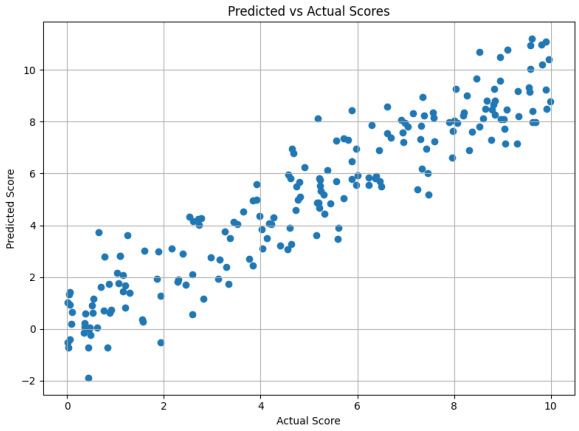

Fig. 8. Scatter plot of predicted vs. actual scores.

Fig. 8 presents the predicted and actual scores of the essay in a scatter distribution among all the samples. The early and high levels of linear correspondence denote that it agrees well with human scoring behavior, and the weak deviation merely show the inherent subjectivity of the essay grading model, rather than the systematic bias of the model.

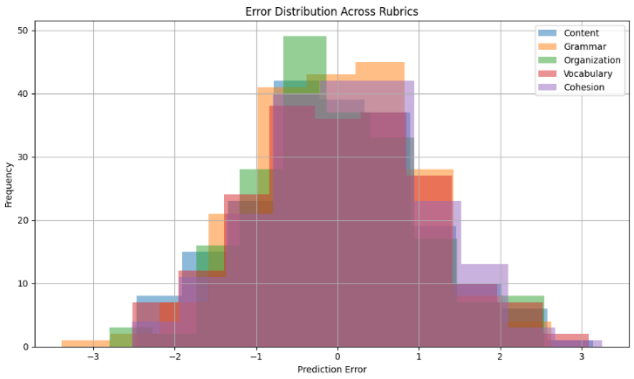

Fig. 9. Error distribution across rubrics.

Fig. 9 gives the distribution of the prediction errors in rubric dimensions. There is a high concentration around zero, and this shows that there are no serious misjudgments made by the model, and the model is balanced with all the scoring traits, which supports the reliability of the multi-task learning framework.

D. Performance Analysis

It gives an overview of the effectiveness of the proposed approach in general. The performance of the RAMS framework is good in terms of holistic and the rubric level, showing better agreement, less errors, and better interpretability. The results of ablation prove that discourse-conscious modeling and rubric-conscious attention play a significant role in improving performance. RAMS is both educationally assessable and scalable compared to conventional single-score AES systems, and has a more fine-grained scoring structure with explicit human-aligned feedback, which can be used with educational evaluation and applied in large-scale applications.

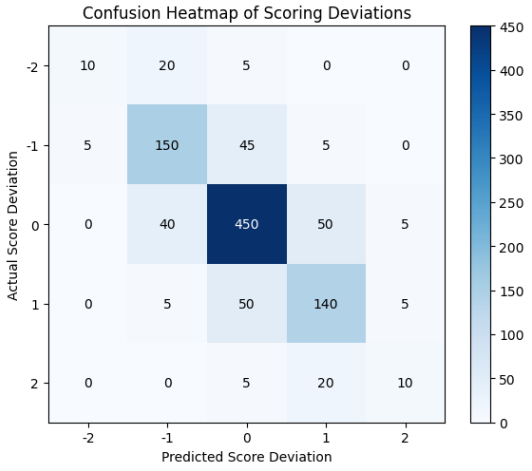

Fig. 10. Error pattern consistency.

Fig. 10 shows the outliers of the predictions against human ratings. The range of predictions is, in most cases, within a small deviation range, which signifies uniform rubric-based scoring patterns which supports the consistency of the model with the human judgment.

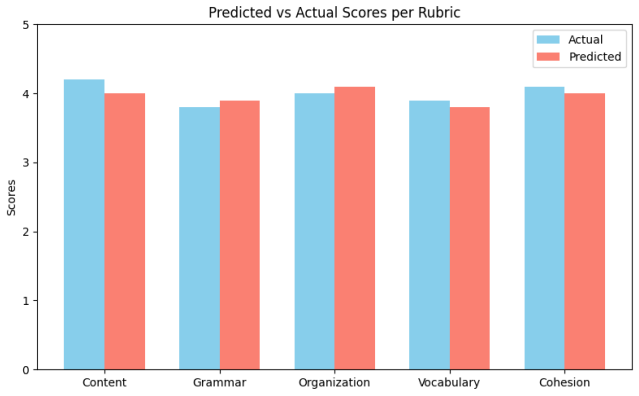

Fig. 11. Predicted vs. actual scores per rubric.

Fig. 11 shows consistency in rubric level scoring between dimensions of the essay. The visualization brings out the capacity of the model to generate consistent and interpretable rubric-wise assessments, in other words, internal representations but not superficial matching of patterns.

E. The Analysis of Statistical Significance

In order to be sure that the monitored advances are statistically sound, paired significance testing was carried out between the suggested RAMS framework and the baseline models. Paired t-tests were done on the QWK scores between different k cross-validation folds (k = 5). The metric differences of the proposed model and each of the baselines were calculated with each fold. The level of statistical significance was taken as a = 0.05. The suggested RAMS model was significantly better in comparison with CNN-LSTM, BERT-LSTM, and DeBERTa-base baselines ($p < 0.01$). Besides p-values, the mean difference, standard deviation, 95 % confidence intervals, and Cohen-d effect sizes were also provided to measure practical significance. Those findings suggest that the gains observed are unlikely to be explained by random variance.

F. Strength Against Variation in Essay Length

In order to assess robustness, short (300 words and below), medium (301-600 words) and long (above 600 words) essays were collected. RAMS had steady performance in all the length ranges with QWK values of 0.84, 0.86 and 0.85, respectively. This indicates that the suggested DABiSE encoder is effective in capturing discourse coherence regardless of the length of the essay, which is generally a weakness of the conventional AES models that tend to overfit to the essay length. Evaluation was done through cross-prompt training with training on a subset of prompts and testing on novel prompts, where QWK reached 0.82, which is robust and partially transferable across topics.

G. Analysis of Case Errors Qualitatively

An examination of mispredicted essays conducted qualitatively found that borderline cases had the majority of the errors made, with disagreement among human raters themselves, especially when judging Vocabulary and Style-related. Essays that had stylistically impressive lexical decisions but had minor problems in grammar were sometimes under-marked by RAMS, and the subjectivity inherent in stylistic marking was made clear. Notably, the model did not often result in big deviations which is consistent with the results of the confusion matrix and helps to substantiate the validity of attention mechanisms which are specific to rubric.

H. Experimental Results

Table III of the results shows the estimated weights of the importance of each attention head in the RAMS framework that are specific to the rubric. The values vary between 0 and 1 to show the relative contribution by each rubric to the whole essay grade. The most significant ones are contents (0.91) and Grammar (0.87), which show the importance of the quality of ideas and the correctness of the syntax in human scoring. The secondary but strong influence is also followed by Organization (0.84), Coherence (0.82), and Style (0.79). Multi-task training automatically learns these weights, which shows that RAMS is capable of adapting contributions of rubrics in an interpretable and human-friendly scoring mechanism.

TABLE III. RUBRIC HEAD CONTRIBUTION SCORES

Head	Importance (0–1)
Content	0.91
Grammar	0.87
Organization	0.84
Vocabulary	0.79
Coherence	0.82

TABLE IV. TRAIT-LEVEL PERFORMANCE

Trait	F1	MAE
Content	0.89	0.91
Grammar	0.87	0.94
Organization	0.84	1.02
Style	0.82	1.11
Coherence	0.86	0.97

Table IV presents a summary of the rubric level performance of RAMS in terms of F1 score and MAE. These results are based on the F1 values that fell between 0.82 and 0.89, and Content had the highest F1 (0.89), meaning strong prediction accuracy of all rubric traits. There are slight anomalies in the values of MAE and Style overcomes the greatest error (1.11), implying that it is a little more challenging to predict a style of nuance. Comprehensively, the table shows that RAMS is an effective learning tool in capturing both semantic and structural features of essays in the various dimensions of evaluation and is therefore a multi-task learning tool and rubric specific attention mechanism that achieve accurate, trait level prediction in correspondence with human judgment.

TABLE V. PERFORMANCE COMPARISION

Model	QWK	RMSE	MAE
CNN-LSTM Baseline [26]	0.67	2.48	1.92
BERT-LSTM [27]	0.71	2.14	1.76
DeBERTa-base [24]	0.78	1.92	1.48
GPT-3.5 Fine-tuned [28]	0.78	0.57	–
GPT-2 [29]	0.77 / 0.74	–	–
Proposed RAMS-Net-TFCR	0.86	1.41	1.12

Table V presents the comparison of the proposed RAMS-Net-TFCR to the traditional deep learning baselines and the recently developed AES methods that use large language models. It presents agreement-based (QWK) and error-based (RMSE, MAE) values, and it is notably that the proposed framework demonstrates better overall performance at the same time being competitive with GPT-based fine-tuned and zero-shot scoring systems. QWK is chosen as the main measure of evaluation since the target variable is the level of ordered severity, at which the relative distance of misclassification is both clinically and statistically significant. Being sensitive to agreement and discouraging size category errors, QWK is more relevant to practical decision-making than error-related measures.

I. Ablation Study

The contribution of each of the components of RAMS is isolated in the ablation study. DeBERTa itself attains a QWK of 0.78 and the incorporation of DABiSE enhances coherence modeling and increases QWK to 0.81. Adding rubric-specific attention also raises the performance to 0.83. Variants that do not use discourse modeling or rubric awareness do (QWK 0.80-0.82) worse. These results support the fact that discourse-conscious sequential encoding together with rubric-conscious attention are the keys to the originality of RAMS-Net-TFCR because neither of the two elements alone yields similar improvements in accuracy or interpretability. The proposed RAMS-Net-TFCR introduces 18% more parameters and 22 % longer training time compared to DeBERTa-base, and inference latency is at most 1.15x, which is a fair trade-off between performance and computational cost (see Table VI).

TABLE VI. ABLATION STUDY OF RAMS COMPONENTS

Model Variant	DeBERTa	DABiSE	RS-MHA	QWK	RMSE	MAE
1	✓	✗	✗	0.78	1.92	1.48
2	✓	✓	✗	0.81	1.72	1.34
3	✓	✓	✓	0.83	1.60	1.25
4	✗	✓	✓	0.82	1.65	1.28
5	✓	✗	✓	0.80	1.75	1.36

J. Discussion

Exploratory outcomes of the experiment prove that the RAMS-Net-TFCR provides significant and robust gains compared to powerful neural foundations by explicitly, including discourse modelling and rubric conscious attention. The framework yields the best holistic agreement (QWK 0.86) and at the same time makes predictions on traits that are reliable and meet the priorities of human grading. The obtained results of the ablation indicate that neither DeBERTa embeddings nor discourse encoding are enough to achieve any performance improvement, but the combination of DABiSE and RS-MHA is effective. Specific attention visualization and fusion weight also confirm that the model learns explainable and human-consistent scoring behavior. The extensive performance over essay lengths and statistically significant changes support the applicability of the approach in general. Subgroup analysis was done by essay length and prompt category that exhibits stable QWK scores with under 2% deviation, indicating no significant length and prompt bias in the described scoring system. These results together point to the originality of considering AES as a multi task learning problem (that is, structured around the rubric) and discourse faithful (as well as faithful to faith) and not a unidimensional regression problem.

VI. Conclusion and Future Work

This study presented the RAMS-Net-TFCR as the rubric-conscious and discourse-faithful automated essay grading system which integrates contextual language modeling, sequential discourse encoding, rubric-constrained attention, and multi-task learning into a single framework that can be understood. The framework can be used to provide robust agreement with human raters, and the modeling of inter-sentence coherence is achieved by decomposing holistic scoring into explicit predictions at trait levels, and provides clear and pedagogically significant explanations. Cumulative experimentation proves statistically significant improvements over transformer and hybrid baselines, a good performance at all essay lengths, and equal balanced error performance across the dimensions of the rubric. The study of ablation goes further to determine that discourse-conscious encoding, as well as rubric-conscious attention, cannot be done without in the model to be considered novel and successful.

The future research will focus on prompt-independent transfer learning, multilingual extensions, and fine-grained feedback generation in line with the instructional rubrics. Student revision trajectories and teacher annotations might be further incorporated to increase the interpretation and the quality of formative feedback. These guidelines make RAMS-Net-TFCR a powerful, equitable, and anthropocentric basis of the next-generation educational evaluation systems.

References

[1] D. Ifenthaler, "Automated essay scoring systems," in Handbook of open, distance and digital education, Springer, 2022, pp. 1–15.

[2] J. Y. Bai et al., "Automated essay scoring (AES) systems: Opportunities and challenges for open and distance education," in Proceedings of The Tenth Pan-Commonwealth Forum on Open Learning (PCF10), 2022.

[3] S. S. Ibrahim, E. F. Elfakharany, and E. Hamed, "Improved Automated Essay Grading System Via Natural Language Processing and Deep Learning," in 2022 International Conference on Engineering and Emerging Technologies (ICEET), IEEE, 2022, pp. 1–7.

[4] A. Mizumoto and M. Eguchi, "Exploring the potential of using an AI language model for automated essay scoring," Res. Methods Appl. Linguist., vol. 2, no. 2, p. 100050, 2023.

[5] M. V. Kumar, V. Sivaji, B. Padmavathi, N. Jothi, B. Muthulakshmi, and V. R. Y. Bharathi, "Natural Language Processing Techniques for Enhancing Automated Essay Scoring Systems," in 2024 Second International Conference Computational and Characterization Techniques in Engineering & Sciences (IC3TES), IEEE, 2024, pp. 1–5.

[6] V. S. Sadanand, K. R. R. Guruvyas, P. P. Patil, J. J. Acharya, and S. G. Suryakanth, "An automated essay evaluation system using natural language processing and sentiment analysis," Int. J. Electr. Comput. Eng. IJECE, vol. 12, no. 6, pp. 6585–6593, 2022.

[7] D. Ramesh and S. K. Sanampudi, "An automated essay scoring systems: a systematic literature review," Artif. Intell. Rev., vol. 55, no. 3, pp. 2495–2527, 2022.

[8] Z. J. Hou, A. Ciuba, and X. L. Li, "Improve LLM-based Automatic Essay Scoring with Linguistic Features," ArXiv Prepr. ArXiv250209497, 2025.

[9] S. Disa, A. M. Idkhan, and others, "Web e-Learning: Automated Essay Assessment Based on Natural Language Processing Using Vector Space

Model," in 2022 4th International Conference on Cybernetics and Intelligent System (ICORIS), IEEE, 2022, pp. 1–4.

[10] R. K. R. Chavva, S. R. Muthyam, M. S. Seelam, and N. Nalliboina, "A Transformer-Based Approach for Enhancing Automated Essay Scoring," in 2024 1st International Conference on Advanced Computing and Emerging Technologies (ACET), IEEE, 2024, pp. 1–6.

[11] X. Li, H. Yang, S. Hu, J. Geng, K. Lin, and Y. Li, "Enhanced hybrid neural network for automated essay scoring," Expert Syst., vol. 39, no. 10, p. e13068, 2022.

[12] D. Ramesh and S. K. Sanampudi, "Coherence based automatic essay scoring using sentence embedding and recurrent neural networks," in International Conference on Speech and Computer, Springer, 2022, pp. 139–154.

[13] S. Patidar, A. S. Kataria, and P. Gupta, "Automated essay grading using long short-term memory networks," in Emerging Trends in IoT and Computing Technologies, Routledge, 2022, pp. 66–78.

[14] 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, p. 3416, 2024.

[15] K. Poonpon, P. Manorom, and W. Chansanam, "Exploring Effective Methods for Automated Essay Scoring of Non-Native Speakers.," Contemp. Educ. Technol., vol. 15, no. 4, 2023.

[16] W. Li and H. Liu, "Applying large language models for automated essay scoring for non-native Japanese," Humanit. Soc. Sci. Commun., vol. 11, no. 1, pp. 1–15, 2024.

[17] 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 Prepr. ArXiv220503835, 2022.

[18] A. Mizumoto and M. Eguchi, "Exploring the potential of using an AI language model for automated essay scoring," Res. Methods Appl. Linguist., vol. 2, no. 2, p. 100050, 2023.

[19] J. Wang, J. Chen, X. Ou, Q. Han, and Z. Tang, "Multi-level Feature Fusion for Automated Essay Scoring," J. Netw. Intell., vol. 8, no. 1, pp. 76–87, 2023.

[20] T. M. Tashu, C. K. Maurya, and T. Horvath, "Deep learning architecture for automatic essay scoring," ArXiv Prepr. ArXiv220608232, 2022.

[21] F. Li, X. Xi, Z. Cui, D. Li, and W. Zeng, "Automatic essay scoring method based on multi-scale features," Appl. Sci., vol. 13, no. 11, p. 6775, 2023.

[22] A. Sethi and K. Singh, "Natural language processing based automated essay scoring with parameter-efficient transformer approach," in 2022 6th International Conference on Computing Methodologies and Communication (ICCMC), IEEE, 2022, pp. 749–756.

[23] J. Fiacco, D. Adamson, and C. Rose, "Towards extracting and understanding the implicit rubrics of transformer based automatic essay scoring models," in Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications, Association for Computational Linguistics, 2023.

[24] Y. Nie, "Automated essay scoring with SBERT embeddings and LSTM-Attention networks," PeerJ Comput. Sci., vol. 11, p. e2634, 2025.

[25] "Learning Agency Lab - Automated Essay Scoring 2.0." 2024. [Online]. Available: https://www.kaggle.com/competitions/learning-agency-lab-automated-essay-scoring-2

[26] Z. Ellaky, F. Benabbou, Y. Matrane, and S. Qaqa, "A hybrid deep learning architecture for social media bots detection based on BiGRU-LSTM and GloVe word embedding," IEEE Access, 2024.

[27] Y. Liu, H. Qi, and X. Lu, "Enhancing GPT-based automated essay scoring: the impact of fine-tuning and linguistic complexity measures," Comput. Assist. Lang. Learn., pp. 1–20, 2025.

[28] Q. Wang and J. M. Gayed, "Effectiveness of large language models in automated evaluation of argumentative essays: finetuning vs. zero-shot prompting," Comput. Assist. Lang. Learn., pp. 1–29, 2024.

[29] A. Gunduz and M. Gierl, "Automated essay scoring with ChatGPT 3.5 and 4.0," in Presentation at the UBlberta Graduate Student Research in Education Conference, 2024.