Minimax Theory and its Applications

Volume 11 (2026), No. 5s, 123-141

Skill-Relational Adaptive Transformer with Proficiency-Conditioned Attention for Context-Aware Multi-Dimensional English Language Assessment

Mr. Yarraguntla Mani Kumar[1], Dr M Surya Bhupal Rao[2], Dr. S. Farhad[3], Dr. Vancha Maheshwar Reddy[4], Dr.R.Jeyalakshmi[5], Elangovan Muniyandy[6], Dr.A.Yaswanth Kumar[7], Maha S. Yaseen[8], Krishnaveni Cherukuri[9]

[1]Research Scholar, Department of English , Koneru Lakshmaiah Education Foundation (Deemed to be University), Greenfields, Vaddeswaram, Guntur, Andhra Pradesh, Email: mkkerenpaul@gmail.com
[2]Associate Professor, Department of CSE (AI&ML) CVR College of Engineering Hyderabad , Email: suryabhupal@gmail.com, Orcid:0000-0002-1096-0799
[3]Associate Professor, Department of English Koneru Lakshmaiah Education Foundation (Deemed to be University) Greenfields Vaddeswaram Guntur Andhra Pradesh , Email: farhad.anu21@gmail.com
[4]Associate Professor, Department of Information Technology, Stanley College of Engineering and Technology for Women Abids, Hyderabad, Telangana, Email: mahesh.vancha@gmail.com
[5]Associate Professor, Dept of MBA,Laki Reddy Bali Reddy college of Engineering (Autonomous), Affiliated to JNTU-K, Mylavaram, NTR District, Andhra Pradesh 521230, Email: pearljeya@gmail.com
[6]Department of Biosciences, Saveetha School of Engineering. Saveetha Institute of Medical and Technical Sciences,Chennai - 602105, Email: muniyandy.e@gmail.com
[7]Associate Professor, Department of Information Technology, R.V.R. & J.C. College of Engineering, Guntur AP, India, Email : *alapatimail@gmail.com*, Orcid id: 0000-0003-1468-7053
[8]Faculty of Arts and Sciences, Department of English Language, Hourani Center for Applied Scientific Research, Al-Ahliyya Amman University, Amman, Jordan, Email: m.yaseen@ammanu.edu.jo, orcid.org/0000-0002-6719-4829
[9]Assistant Professor, Department of CSE-Data Science, R.V.R. & J.C. College Of Engineering, Chowdavaram, Guntur, Email: cherukuriveni@gmail.com

Abstract
Automated essay assessment enables scalable and timely evaluation of English learners' writing, where quality spans cohesion, syntax, vocabulary, phraseology, grammar, and conventions. Reliable multidimensional scoring remains difficult because essay quality depends on writing context and on interdependencies among analytic traits. Existing trait-aware models often underrepresent topic-sensitive dimensions, compress token-level information, and rarely combine proficiency signals with explicit skill relationships. This study proposes the Skill-Relational Adaptive Transformer (SRAT) for context-aware prediction of six analytic scores and one holistic score. It refines essay representations using estimated proficiency. It also integrates writing-topic context adaptively and models dependencies among writing skills. The publicly available ELLIPSE corpus, containing 3,911 training and 2,571 testing learner essays with 44 topic labels and seven continuous dimensions, was used. A shared-weight DeBERTa-v3-base encoder independently represents essays and writing contexts. An auxiliary branch estimates preliminary proficiency, which conditions an additional multi-head attention module through token-pair biases, while an adaptive gate fuse refined essay and projected context features. Six skill representations are refined through a learned inter-skill relation matrix, and separate regression heads predict analytic and holistic scores. Performance was evaluated on the official testing partition using QWK, RMSE, MAE, and PCC, complemented by ablation and computational-cost analyses. Every component improved performance, with writing-context encoding contributing most, and RMSE fell 23.2% relative to the backbone, while the complete model added only 8.6 M parameters and 3.3 ms latency. SRAT achieved QWK 0.912, RMSE 0.258, MAE 0.206, and PCC 0.921, exceeding the reported results of compared methods.
Keywords: Automated essay scoring; Multidimensional writing assessment; Proficiency-conditioned attention; Skill-relational modeling; Gated context integration; ELLIPSE corpus

1. Introduction

Automated essay assessment has become an important research area in educational technology because it can support timely, consistent, and scalable evaluation of learners' writing[1], [2]. English writing ability is complex; quality is based not only on what is written, but also on such factors as cohesion, syntax, vocabulary, phraseology, grammar, and writing conventions[3]. These dimensions are interconnected; however, each

dimension conveys a unique aspect of writing proficiency and develops at different rates among students. Thus, providing one holistic mark does not fully cover a student's strong and weak sides, while evaluating each dimension separately may fail to recognize the interrelation between writing skills[4]. Automated assessment thus necessitates the use of models which incorporate the meaning, language, context of assessment, and interdependence between multiple assessment dimensions[5]. Transformer-based language models provide contextual representations of text, creating opportunities for more detailed scoring[6]. However, effective assessment also depends on how essay content is interpreted in relation to its writing topic and how contextual evidence is translated into coherent, dimension-specific predictions[7].

Recent automated essay-scoring research has increasingly moved beyond holistic grading toward joint prediction of overall quality and individual writing traits, using dual-scale BERT representations and multi-task learning to capture relationships between assessment levels[8]. Hybrid architecture approaches have also fused convolutional and Transformer components with prompt-conditioned and learner-conditioned representations to facilitate nuanced trait scoring and rubric-based feedback, but their resource-intensive nature and reliance on learner data is problematic[9]. Other approaches have combined recurrent, Transformer, and gradient boosting elements to capture linguistic and semantic aspects of essays in the English language scoring[10]. Large language models are promising for evaluating students' science writing with good alignment with grades provided by instructors based on a rubric evaluation system[11]. Prompt-based studies further indicate that anchor essays can improve agreement between LLM and human scores [12]. Nevertheless, analytic scoring remains affected by rater variability, score compression, and difficulty evaluating nuanced writing qualities [13]. This research emphasizes the importance of integrating context-based representations along with consistent predictions for various aspects of writing.

To address these challenges, this study proposes SRAT, a context-aware framework for predicting six analytic writing scores and one holistic score. The framework employs the ELLIPSE dataset, keeping the text of the essay and related information about the topic as separate inputs to the model and the seven assessment dimensions as separate prediction targets. The shared-weight DeBERTa-v3-base encoder produces context-based representations of both essays and writing contexts independently. An auxiliary prediction branch is used to estimate preliminary essay proficiency and then passed as a conditioning signal to an extra multi-head attention module to further refine the token-level representations of the essay. At the same time, a compatible feature space is generated to represent the writing-context and then a combined representation is sent to the adaptive gate which combines it with the refined essay representation. These are then linked to each other using a learned inter-skill relation matrix and relation refinement to build a matrix of inter-skill relations, and a separate regression branch estimates holistic quality. Experiments explore predictive agreement and error, analyse the contribution of individual components via ablations and analyse the computational requirements to evaluate the effectiveness and implementation cost of the framework.

1.1 Problem statement

Reliable multidimensional English essay assessment remains challenging because quality depends on writing context, prompt variation, and relationships among analytic traits. Existing systems for scoring that take into account trait-based scores will fail to account for prompt-based dimensions since the prompt material will not be used as external information[14]. Continuous adaptation to new prompts that may be presented is also an aspect that remains unexplored[15]. Moreover, more research is needed on generalization for narrative, reflective, and creative domains, while computational challenges hinder practical implementation[16]. To solve these deficiencies, an architecture independent of the available writing topic, accounting of the token-level attention conditioned by the predicted proficiency and the relationship between six analytical skills, along with the overall result, is built. Continual learning and cross-genre generalization should be assessed individually.

1.2 Research Motivation

The study is motivated by the need for automated assessment that considers both what learners write and the topic they address. The previous approaches can be limited to teaching traits of writing, may embed information at a token level, or may provide weak linkage between contextual and proficiency indicators. Thus, a common representation is required to support better refinement of essay representations based on estimated proficiency; to incorporate writing context adaptively, and to learn dependency among analytic skills to support more coherent and informative multidimensional representations in varied writing contexts and assessment settings.

1.3 Significance of the Study

The study brings together a comprehensive representation of the content of an essay, of the context of its writing and of the relationships between the analytic skills. It is designed with seven outputs so it can offer a comprehensive assessment with specific feedback that may assist teachers to identify instruction areas that need support. Ablation and computational analyses complements evaluation on ELLIPSE to give evidence on component contributions, predictive performance and architectural cost. These analyses can be used to guide future work in building automated assessment systems for the ELL and educational feedback that are context-aware and that can be evaluated transparently.

1.4 Key Contributions

- SRAT, a context-aware framework, jointly predicts six analytic writing scores and one holistic score from essay text and writing topic.

- The ELLIPSE corpus, comprising 3,911 training and 2,571 testing essays with seven continuous dimensions, supplies separate essay and topic inputs.
- A proficiency-conditioned attention module injects an auxiliary proficiency estimate into token-pair biases, refining essay representations beyond the shared DeBERTa-v3-base encoder.
- Gated fusion combines essay and projected topic features, while a learned inter-skill relation matrix models dependencies among six analytic dimensions.
- SRAT achieves QWK 0.912, RMSE 0.258, MAE 0.206, and PCC 0.921, outperforming compared methods with only 8.6 million additional parameters.

1.5 Organization of the Remaining Sections

The rest of this study is structured as follows: The related studies reviewed in Section 2 include those pertaining to automated essay assessment and multidimensional scoring. Section 3 describes SRAT, covering its dataset preparation, contextual encoding, attention based on proficiency, gated integration, skill-relational modeling, and score prediction. Section 4 reports the experimental setup and discusses dataset characteristics, model behavior, efficiency, ablation, and comparative performance. Section 5 concludes the study and outlines future research.

2. Related Works

Sun et al. [14]introduced a new multi-trait essay scoring method called UniMTES which integrates both trait-description representations and bidirectional attention to link the essay text with the criteria. The experiments on ASAP and ASAP++ showed that the scoring improved, with an average improvement of 3.3% in QWK and increased parameter and runtime efficiency. Original prompt content is not, however, used for external knowledge. This prevents the evaluation of prompt-sensitive traits, and weaker Content and Prompt Adherence effects than the baseline were seen for Prompt 6, reflecting the importance of the contextual information.

Han et al. [15]developed a multi-level disentanglement framework for fine-grained multi-prompt essay scoring to address semantic variations in writing traits across prompts. The approach distinguishes the meaning of the prompt, while keeping representations of traits relevant to the target task. Comparing with seven automatic essay scoring approaches and LLM-based zero/few-shot methods revealed benefits in scoring efficiency, in terms of reduced memory usage and lower latency.However, the extent to which continuous adaptation is possible to the new prompts introduced was not explored, and the ability of the framework to cope with changing assessment tasks after significant retraining was not tested.

Ren et al. [16]introduced the adversarial training method with topic-independent representations and topic-specific representations, structural features, and attention mechanisms to tackle the cross-topic English essay scoring problem. The average cross-prompt QWK was 0.856, which showed good performance for the writing in different writing topics. But its applicability to various styles of writing, such as narrative, reflective, and creative writing is not well studied. There are important aspects of deployment in educational contexts of limited resources, however, that are computationally problematic as well as the desire for more genre adaptability and computational efficiency in deployment.

Song et al. [17]proposed an external-knowledge similarity-transfer framework to improve cross-prompt essay trait scoring by enriching shared representations and aligning task-specific traits. It develops further representations of knowledge for the task via additional extra-essay routes, and transfers information about the similarity of samples from source to target prompts. Experiments with the ASAP++ indicated that it is superior to the traditional method. Some parts of the performance, however, are still subject to prompt bias and substantial domain differences, and differences in task requirements and scoring scales pose an obstacle to criterion transfer. In addition, the framework is underdeveloped, as there are topics for which there is little data.

Miao andXu[18]proposed the multi-scale key-information weighting method with BERT for automated essay scoring. The method identifies the important information and weights at the document, paragraph, sentence and token level, as well as data augmentation. This was demonstrated using Evaluation on the ASAP dataset, where the improved holistic scoring performance was demonstrated. However, the approach is only applied to overall essay scoring, and has not yet been developed for the grading of individual traits; the weighting and augmentation strategy is yet to be solved for individual trait scores. Also, it has the less preferred prompt-similarity based weighting for cross-prompt scoring tasks.

Lohmann et al.[19]examined the accuracy of linguistic-feature models, DistilBERT contextual embeddings, and hybrid models for analyzing content, organizational and language features of L1 and L2 English essays. Using the four prompts from the ASAP and MEWS datasets, the study found that there was evidence that linguistic and contextual information can be used together – the hybrid models performed better when both types of information were used together than single-resource models. However, embedding contextualized representations at the essay level sacrifices token-level information and fine-tuning the Transformer model for the specific task in a trait-sensitive manner has not been widely studied, as not enough data exists yet.

Tang et al.[20]examined the accuracy of LLMs for multi-dimensional evaluation of writing and how accurate they are compared to human judgments. The study focused on prompt engineering and temperature settings especially the dimensions of Ideas and Organization. Based on the results, the accuracy of GPT-4's scores was different in various areas of the rating, with higher accuracy in Ideas and Organization compared to Conventions. This imbalance in the dimension level performance suggests that just as with general-purpose

language models, it is difficult to assess linguistic conventions and needs to be improved on for comprehensive multidimensional assessment of writing.

Pack et al.[21]investigated the validity and reliability of LLM for automating the scoring of ELL writing. The scientists then tested this by presenting 119 essays from an English level placement test to the AI models, and human raters, for comparison. To test this, the researchers used an English language placement test to choose 119 essays and sent them to the AI models and human ratings. The results suggest that certain models have the potential to be applied in some contexts such as second-rater support, human-rating disagreement, and low-stakes formative assessment. The findings should also raise awareness, however, about the need to investigate aspects of validity, reliability and ethical issues before the technology is introduced in school in freedom.

The related studies reveal limitations in jointly modeling writing context, proficiency variation, token-level linguistic information, and interdependencies among multiple writing traits. Current methods tend to focus on prompt-specific representations or on holistic scoring or on isolated evaluation of traits, but fail to combine these complementary assessment dimensions. To tackle this, SRAT fuses shared-weight DeBERTa-v3 essay and context encoding, proficiency-conditioned multi-head attention, gated feature integration, and skill-relational modeling. It is a unified architecture that provides context-aware learning and integrated prediction of seven analytic dimensions, in addition to the overall essay assessment.

3. Proposed Skill-Relational Adaptive Transformer Framework

The proposed SRAT framework aims to conduct context-aware multi-dimensional evaluation of English essays by combining essay representation learning and modeling of the relations between writing skills. The methodology includes the independent encoding of the essay and writing context, the proficiency-conditioned attention, the adaptive integration of features, and the representation of relational skills in a common assessment architecture. The essays, writing-context information, and seven continuous assessment dimensions that are needed for the suggested task are available in the ELLIPSE data set, which is publicly available. First, the essay text and writing-context text are preprocessed separately and then tokenized, and then a shared-weight DeBERTa-v3-base encoder is used to process them. The resulting contextual representations run on separate processing paths and enable the estimation of essay proficiency and projection of features in writing contexts without mixing up the original order of the tokens. Next, the estimated proficiencies serve as the basis for an additional multi-head attention mechanism which further improves essay representations at the token level. A gated integration model integrates the refined essay representation and the features extracted from the text of the writing. The relationally modeled analytic dimensions are also integrated, along with a holistic prediction, using dedicated regression heads and providing an integrated 7-dimensional assessment output.

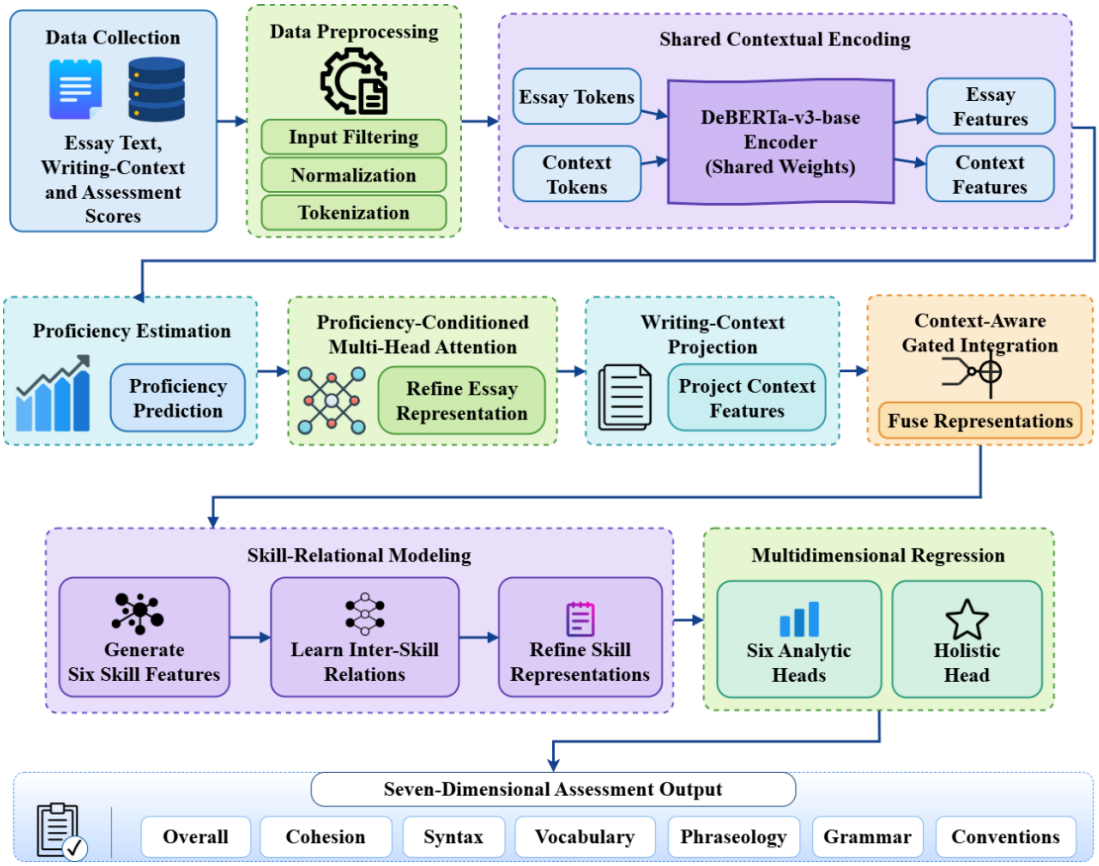

Figure 1. Proposed SRAT Framework for Multidimensional Essay Assessment

Figure 1 shows the proposed SRAT framework for multidimensional essay assessment. Both the essay text and the writing-context information are first preprocessed and then fed into the shared DeBERTa-v3-base LLM encoder.The essay text and the writing-context information are separately preprocessed, and then separately input into the shared DeBERTa-v3-base LLM encoder. The essay pathway estimates proficiency and fine tunes token representations, while the writing-context pathway projects topic features. Gated integration uses both representations to enable skill-relational modelling and provides separate analytic and holistic regression branches to produce 6 skill scores and 1 overall score.

3.1 Data Collection

The publicly available ELLIPSE dataset was collected from its GitHub repository[22], which provides English learner essays, associated writing-context information, and multidimensional assessment annotations. Each record in the corpus has 26 attributes, and a total of 3,911 training records and 2,571 testing records. The full_text element contains the essay text and the prompt element contains the related writing prompt. There are seven dimensions of continuous assessment in the dataset: Overall, Cohesion, Syntax, Vocabulary, Phraseology, Grammar, and Conventions. These indicators offer the dimensions of the target variables to be assessed in a multidimensional essay. The 44 different topic labels correspond to the available writing-context information. The proposed framework is based on the separation of essay text and associated writing-context information for their learning from context and evaluation.

3.2 Data Preprocessing

The preprocessing stage prepares essay content and associated writing-context information for independent transformer encoding while preserving linguistic characteristics relevant to writing assessment. The procedure includes input selection, text normalization, and separate tokenization.

3.2.1 Input Selection and Attribute Filtering

The dataset is filtered to retain essay text (full_text), topic-based writing context, and seven continuous assessment targets. Record identifiers and demographic attributes, constant task categories and irrelevant metadata are filtered out, leaving valid textual inputs for SRAT, while the individual assessment dimensions for multidimensional score prediction remain.

3.2.2 Text Normalization and Tokenization

Texts that are in the form of an essay or writing-context are normalized conservatively, but with punctuation, capitalization, sentence boundaries and word order retained. The DeBERTa-compatible tokenizer is used to process both the input and output, yielding token sequences and attention masks for both and keeping those differences in the proposed architecture for the representation learning and contextual processing of both inputs. The tokenization process is represented in eqn. (1).

$$X_e = \mathrm{Tok}(T_e), \qquad X_c = \mathrm{Tok}(T_c) \tag{1}$$

Here, $T_e$denotes original essay text,T_c denotes the associated writing-context text. $\mathrm{Tok}(\cdot)$is the tokenizer operation that converts textual inputs into subword tokens and corresponding attention masks. X_e and X_c denote the separately tokenized essay and writing-context sequences, respectively.

3.3 Shared Contextual Encoding

The tokenized essay and writing-context sequences are processed independently through one shared-weight DeBERTa-v3-base encoder. This arrangement establishes a common contextual feature space while retaining separate pathways for essay content and associated writing-context information.

3.3.1 Shared-Weight DeBERTa-v3-base Encoder

The proposed architecture uses DeBERTa-v3-base as the only pretrained language backbone for contextual representation learning. It has parameters common to both essay- and writing-context-training, which are processed forward separately. As a result, the encoder is able to generate the contextual representations with a coherent semantic control without having to concatenate the two input sequences. This shared encoding operation is formulated in eqn. (2).

$$H_e = F_\theta(X_e), \qquad H_c = F_\theta(X_c) \tag{2}$$

Here, $F_\theta(\cdot)$ denotes the DeBERTa-v3-base contextual encoder parameterized by shared weights θ. H_e represents the contextual token features extracted from the essay sequence, whereas H_c represents the contextual token features extracted independently from the writing-context sequence. X_e and X_c denote their respective tokenized inputs.

3.3.2 Essay and Writing-Context Feature Extraction

The encoder generates token-level contextual information containing linguistic and semantic features from the sets of tokens from each of the input sequences. These token matrices are then transformed into a compact representation by a pooling operation, and then used to estimate the proficiency and project the writing context. The independent feature extraction process is expressed in eqn. (3).

$$h_e = \mathrm{Pool}(H_e), \qquad e_c = \mathrm{Pool}(H_c) \tag{3}$$

Here, h_e denotes the pooled essay representation, whereas e_c represents the pooled writing-context representation. H_e and H_c denote their corresponding contextual token matrices. $\mathrm{Pool}(\cdot)$ represents the masked pooling operation applied independently to each sequence.

3.4 Provisional Proficiency Estimation

The pooled essay representation is used to give an initial representation of overall writing ability prior to the attention refinement stage. This estimate provides an essay-based conditioning signal which can impact on the attention assigned to the subsequent token without using assessment values of the ground truth.

3.4.1 Auxiliary Proficiency Prediction

A regression transformation of the auxiliary kind converts the pooled essay representation into a preliminary holistic score.This prediction is based purely on features extracted from the essays, and writing-context information is not used directly to determine the initial estimate of proficiency level. The estimation operation is described in eqn. (4).

$$\hat{p} = f_p(h_e) = W_p h_e + b_p \tag{4}$$

Here, $\hat{p}$ denotes the predicted preliminary overall proficiency score, whereas $f_p(\cdot)$ represents the auxiliary regression function. W_p and b_p denote its transformation weight matrix and bias vector, respectively. h_e represents the pooled essay feature vector supplied to the proficiency estimation component.

3.4.2 Proficiency Conditioning Representation

The preliminary estimate of proficiency is converted to a conditioning vector that can be used in token-pair attention calculations. This change allows the next attention mechanism to use essay-level proficiencies to make relationships between individual contextual tokens. The conditioning representation is generated using eqn. (5).

$$c_p = \mathrm{GELU}\left(W_{cp}\hat{p} + b_{cp}\right) \tag{5}$$

Here, c_p denotes the learned proficiency-conditioning vector, whereas W_{cp} and b_{cp} represent the corresponding trainable projection parameters. $\hat{p}$ represents the preliminary proficiency prediction, and $\mathrm{GELU}(\cdot)$ denotes the Gaussian Error Linear Unit activation function.

3.5 Proficiency-Conditioned Multi-Head Attention

The added attention mechanism enhances essay token representations by adding the estimated proficiency signal to token-pair interaction scores. It is an additional component that explicitly refines the representation of the encoder using self-attention mechanism in a way dependent on the proficiency level.

3.5.1 Query, Key, and Value Transformation

The query, key, and value representations for the various attention heads are linear transformed independently from the contextual essay token matrix. The projections enable individual heads to see different patterns of token-interaction, with the same essay-level terms of proficiency conditioning information. The transformations are defined in eqn. (6).

$$\begin{aligned} Q^{(m)} &= H_e W_Q^{(m)}, \\ K^{(m)} &= H_e W_K^{(m)}, \\ V^{(m)} &= H_e W_V^{(m)} \end{aligned} \tag{6}$$

Here, $Q^{(m)}$, $K^{(m)}$, and $V^{(m)}$ denote the query, key, and value matrices for attention head m. $W_Q^{(m)}$, $W_K^{(m)}$, and $W_V^{(m)}$ represent their corresponding trainable projection matrices. H_e denotes the contextual essay token matrix supplied to the additional attention module.

3.5.2 Proficiency-Dependent Attention Bias

A proficiency-dependent interaction term is calculated for each query-key token pair within every attention head. This is a formulation that allows to account for the proficiency estimated separately for each token-pair score, instead of using the same scalar adjustment for all the attention positions for a single score. The conditioning bias is formulated in eqn. (7).

$$B_{ij}^{(m)} = w_m^T \tanh\left(W_{qm} q_i^{(m)} + W_{km} k_j^{(m)} + W_{pm} c_p + b_m\right) \tag{7}$$

Here, $B_{ij}^{(m)}$ denotes the proficiency-dependent attention bias between query token i and key token j for attention head m. $q_i^{(m)}$ and $k_j^{(m)}$ represent their projected query and key vectors. W_{qm}, W_{km}, and W_{pm} denote trainable transformation matrices, whereas w_m and b_m represent the output weight vector and bias. c_p supplies the estimated proficiency-conditioning vector.

3.5.3 Attention Weight Calculation

The features at the token level are then refined to create a summary of the essay, which is then projected onto the writing-context. This process allows the storage of the skilled information and condenses the tokens into one feature vector. Attention weights are normalized using eqn. (8).

$$A^{(m)} = \mathrm{softmax}\left(\frac{Q^{(m)} K^{(m)T}}{\sqrt{d_k}} + B^{(m)} + M_{\mathrm{pad}}\right) \tag{8}$$

Where, $A^{(m)}$represents the normalized attention matrix for head m, d_k represents the dimensionality of the projected key vectors. $B^{(m)}$ contains the proficiency-dependent token-pair biases, and $M_{\mathrm{pad}}$is the padding mask. $Q^{(m)}$ and $K^{(m)}$are the corresponding projected query and key matrices.

3.5.4 Proficiency-Refined Essay Representation

The normalized attention weights combine value vectors and output the enhanced contextual features for each head of attention. The outputs are concatenated and fed to a projection layer to produce an essay token representation with improved proficiency.The refinement operation is expressed in eqn. (9).

$$Z_p = \text{Concat}\left(A^{(1)}V^{(1)}, \ldots, A^{(n_h)}V^{(n_h)}\right)W_O \quad (9)$$

Here, $Z_p$is the proficiency-refined essay token representation, n_h represents the number of attention heads. $A^{(m)}$ and $V^{(m)}$represent the normalized attention and value matrices for head m, respectively. $\text{Concat}(\cdot)$ combines the individual head outputs, and $W_O$denotes trainable output projection matrix.The token-based refined feature representations are then pooled to form an essay representation which is able to be used for the projection of the writing context. This process retains the proficiency-specific information while condensing the tokens into a single feature vector. Pooling is performed using eqn. (10).

$$Z_{\text{pool}} = \text{Pool}\left(Z_p\right) \quad (10)$$

Here, $Z_{\text{pool}}$is the pooled proficiency-refined essay representation, $Z_p$denotes the token-level output of the proficiency-conditioned attention module. $\text{Pool}(\cdot)$ denotes the masked pooling operation applied to the refined token representations.

3.6 Writing-Context Feature Projection

The projection and activation of the writing-context representation is trained such that it resembles the fine-grained essay representation.This transformation retains contextual information, yet keeps it separate from features derived from the essay for later gated integration to yield a comparable representation feature vector for combined representation learning. The projection process is defined in eqn. (11).

$$e_c^* = \text{GELU}(W_{cc}e_c + b_{cc}) \quad (11)$$

Here, e_c^* denotes the projected writing-context representation, whereas e_c represents the independently pooled context feature vector. W_{cc} and b_{cc} denote the trainable projection matrix and bias vector, respectively. $\text{GELU}(\cdot)$ introduces nonlinear transformation into the projected writing-context representation.

3.7 Context-Aware Gated Integration

An adaptive gating mechanism is proposed to combine the refined essay representation and the projected writing-context representation. This component specifies the relative impact of essay-based and writing context-based information in the common multi-dimensional assessment representation.

3.7.1 Adaptive Gate Generation

Feature-wise integration weights are computed using the pooled essay representation of the proficiencies and the projected essay vector for the context of writing. These weights are restricted to the [0,1] interval, allowing the selection of information from both independent representation pathways in an adaptive way.The gating operation is expressed in eqn. (12).

$$g = \sigma\left(W_g\left[Z_{\text{pool}} \| e_c^*\right] + b_g\right) \quad (12)$$

Here, g denotes the learned gating vector, $Z_{\text{pool}}$is the pooled proficiency-refined essay features. e_c^*represents the projected writing-context representation, and $\|$ indicates vector concatenation. W_g and b_g represent the trainable gate parameters, while $\sigma(\cdot)$ denotes the sigmoid activation function.

3.7.2 Context-Aware Feature Fusion

The generated gate determines the relative amount of refined essay information and projected writing-context information that are used for feature integration. Combining the features in a feature-wise weighted combination results in a unified representation with essay-specific properties as well as information relevant to the writing-context. Formulate an integration operation into eqn. (13).

$$h_c = g \odot Z_{\text{pool}} + (1 - g) \odot e_c^* \quad (13)$$

Here, $h_c$represents the resulting shared context-aware representation, $g$denotes learned feature-wise gate. Z_{pool} and $e_c^*$are the pooled refined essay and projected writing-context vectors, respectively. The symbol $\odot$ indicates element-wise multiplication between compatible feature vectors.

3.8 Skill-Relational Modeling

Prior to learning relational interactions, the integrated representation of the context is transformed to six separate analytic skill representations. This component captures the dependency between the dimensions of the writing, without losing the distinct feature representations for the assessment of each dimension.

3.8.1 Analytic Skill Representation Generation

The six independent transformation functions produce feature representations that represent the notions of Cohesion, Syntax, Vocabulary, Phraseology, Grammar, and Conventions. For each transformation, the shared context-aware representation is shared, and dimension-specific characteristics relevant to the transformation's corresponding writing skill are learned. The six representations are assembled according to eqn. (14).

$$S = \begin{bmatrix} s_1 \\ s_2 \\ s_3 \\ s_4 \\ s_5 \\ s_6 \end{bmatrix}, \qquad s_i = \text{GELU}(W_i h_c + b_i) \quad (14)$$

Here, S denotes the six-skill representation matrix, whereas s_i represents the feature vector for analytic dimension i. W_i and b_i denote dimension-specific transformation parameters, respectively. h_c represents the shared context-aware representation supplied to all six skill transformations.

3.8.2 Inter-Skill Relation Learning

The six analytic skill representations that are extracted from the integrated essay features are used to represent the dependencies between these skills in a trainable relation matrix. Row-wise normalization transforms the relation scores to attention distributions on the modeled dimensions of writing, allowing the relational component to discern the relative importance of connected skills.The normalized relation matrix is calculated using eqn. (15).

$$R_a = \text{softmax}(R,\text{row-wise}) \quad (15)$$

Here, R_a denotes the normalized inter-skill relation matrix, whereas R represents the trainable relation-score matrix. Each row of R_a describes the learned relative relationships between one analytic skill and the remaining dimensions. The row-wise softmax function guarantees that the relations' weights in each row follow a normalized distribution.

3.8.3 Relational Skill Feature Refinement

Learned relation weights combine information from the analytic skill representations that are connected preceding dimension-specific regression. The initial skill representations are stored in a residual connection and the resulting relational features are stabilized by a layer norm. The relational refinement operation is specified in the following eqn. (16).

$$S_{\text{rel}} = \text{LayerNorm}(S + R_a S W_s) \quad (16)$$

Here, S_{rel}represents the refined relational skill matrix, Sdenotes initial six-skill feature matrix. R_a contains the normalized inter-skill relation weights, and $W_s$is the trainable relation transformation matrix. $\text{LayerNorm}(\cdot)$ represents layer normalization applied to the residual relational features.

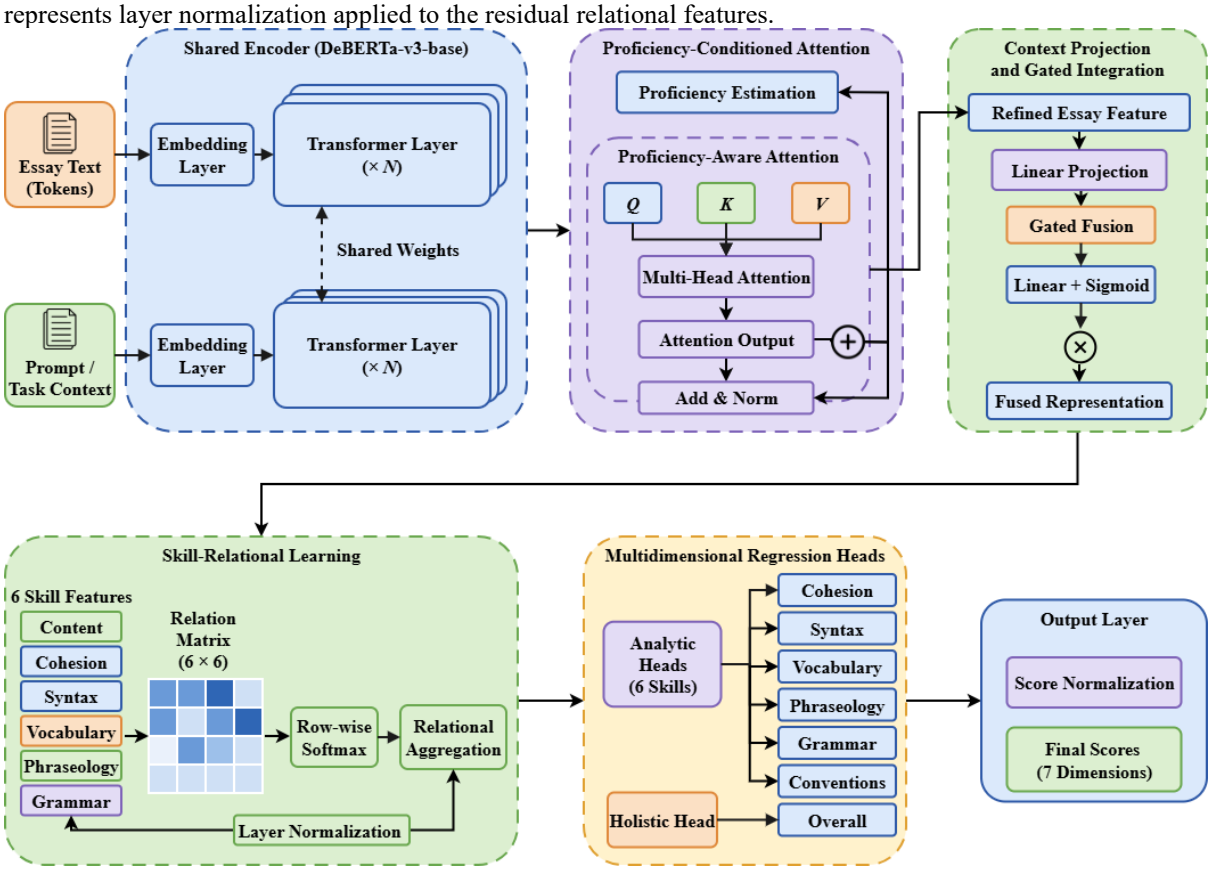

Figure 2. Proposed SRAT Architecture for Multidimensional Essay Assessment

Figure 2 illustrates the architecture of the proposed SRAT model, integrating shared DeBERTa-v3 encoding, proficiency-conditioned multi-head attention, and context-aware gated integration. The fused representation is fed into the skill-relational module where the six analytic writing dimensions are processed in terms of feature transformation, relation learning and relational aggregation. Finally, multiscore regression heads produce seven assessment scores including overall and six analytic writing dimension scores in multidimensional essay assessment.

3.9 Multidimensional Regression

The refined analytic skill representations and shared context-aware representation are supplied to separate regression pathways. Six analytic predictions are generated from relational skill features, whereas the holistic prediction is derived directly from the integrated representation.

3.9.1 Analytic Score Prediction

Each relational skill representation is processed by its corresponding regression head to generate a continuous assessment score for the associated writing dimension. The six analytic predictions preserve dimension-specific information while reflecting the relationships learned through the skill-relational modeling component. Their prediction process is represented in eqn. (17).

$$\hat{\mathbf{y}}_a = \left[f_1(s_{\text{rel},1}), f_2(s_{\text{rel},2}), \dots, f_6(s_{\text{rel},6})\right]^T \quad (17)$$

Here, $\hat{\mathbf{y}}_a$ denotes the six-dimensional analytic prediction vector, whereas $s_{\text{rel},i}$ represents the refined relational representation of analytic dimension i. $f_i(\cdot)$is the corresponding dimension-specific regression function, producing predictions for Cohesion, Syntax, Vocabulary, Phraseology, Grammar, and Conventions.

3.9.2 Holistic Score Prediction

The six relational analytic regression heads work independently on the shared context-aware representation along the holistic assessment pathway. This pathway uses the integrated essay and writing-context features to make an overall judgment of essay quality. The overall prediction is given by eqn. (18).

$$\hat{y}_{\text{Overall}} = f_o(h_c) \quad (18)$$

Here, $\hat{y}_{\text{Overall}}$is the predicted holistic assessment score, $f_o(\cdot)$is the holistic regression function. $h_c$represents shared context-aware representation supplied directly to the holistic prediction head

3.9.3 Final Assessment Representation

The holistic and analytic predictions are combined into a seven-dimensional output vector, where Overall is first, and the six dimensions of analytic writing are thereafter. The proposed SRAT framework produces the total assessment output consistently, with this order for the representation. The final output is determined in eqn. (19).

$$\hat{\mathbf{y}} = \left[\hat{y}_{\text{Overall}}, \hat{y}_{\text{Cohesion}}, \hat{y}_{\text{Syntax}}, \hat{y}_{\text{Vocabulary}}, \hat{y}_{\text{Phraseology}}, \hat{y}_{\text{Grammar}}, \hat{y}_{\text{Conventions}}\right]^T \quad (19)$$

Here, $\hat{\mathbf{y}}$ denotes the final seven-dimensional assessment vector. $\hat{y}_{\text{Overall}}$ represents the holistic assessment prediction, whereas the remaining six variables represent the predicted analytic scores for Cohesion, Syntax, Vocabulary, Phraseology, Grammar, and Conventions, respectively. These seven outputs collectively represent the multidimensional assessment generated by the proposed SRAT framework.

Algorithm 1. Proposed SRAT-Based Multidimensional Essay Assessment

Input:
Essay texts and associated writing-context texts
Output:
Seven-dimensional assessment vectors
Initialize the shared DeBERTa-v3-base encoder
For each essay record do
If essay text or writing-context text is unavailable then
Skip the current record
End if
Normalize essay text and writing-context text independently
Tokenize both inputs into separate token sequences
Extract contextual features using the shared encoder
Pool essay features and writing-context features separately
Estimate provisional proficiency from pooled essay features
Transform estimated proficiency into a conditioning vector
For each attention head do
Generate query, key, and value representations
For each query-key token pair do
Calculate proficiency-dependent attention bias
Incorporate bias into scaled dot-product attention scores
End for
Normalize attention scores using softmax
Aggregate value representations using attention weights
End for
Concatenate attention-head outputs and apply output projection
Pool proficiency-refined essay token representations
Project independently extracted writing-context features
Calculate feature-wise adaptive gating weights
Fuse refined essay features with projected writing-context features
For each of the six analytic writing dimensions do
Generate the corresponding skill-specific representation
End for
Construct the inter-skill relation matrix
Normalize inter-skill relationships using row-wise softmax

```
    Refine skill representations through relational aggregation
    For each relationally refined skill representation do
      Generate the corresponding analytic assessment score
    End for
    Predict holistic assessment score from the fused representation
    Assemble Overall first, followed by six analytic scores
    Store the seven-dimensional assessment vector
  End for
  Return all seven-dimensional assessment vectors
```

Algorithm 1 presents the entire process of the proposed SRAT framework from independent preprocessing of essay and writing-context inputs to the inference results. That are represented separately in a shared DeBERTa-v3-base encoder. The multi-head attention and the integration of the features that are expected to be proficient at the token level are handled by a token-level multi-head attention, whereas the features that are expected to be proficient at the writing context level are integrated using adaptive gating. The fused representation leads to six relationally refined analytic predictions and one holistic prediction that are brought together into the final seven-dimensional assessment output for each essay record.

4. Results and Discussion

This section presents the experimental evaluation of the proposedSRAT framework on the ELLIPSE dataset. The results are presented in the order of the methodology: dataset characteristics and outcomes of preprocessing, shared contextual encoding, proficiency estimation, proficiency-conditioned attention, gated integration, skill-relational modeling, and multidimensional prediction. Training behavior and computational efficiency is then studied and followed by an ablation study and performance comparison with previous techniques. Table 1. Evaluation is based on the official training and testing partitions with the aim of fair, reproducible and consistent reporting all across experiments.

Table 1. Experimental setup and simulation parameters used for SRAT

Parameter	**Value**
Framework	PyTorch 2.1, Hugging Face Transformers
Hardware	NVIDIA A100 GPU (40 GB)
Backbone encoder	DeBERTa-v3-base (shared weights)
Maximum essay length	512 tokens
Maximum context length	32 tokens
Hidden dimension	768
Attention heads (additional module)	8
Optimizer	AdamW
Learning rate	2×10^{-5} (encoder), 1×10^{-4} (task layers)
Batch size	16
Epochs / early stopping	15 / patience 3
Warm-up ratio	0.1
Weight decay / dropout	0.01 / 0.1
Loss function	Smooth L1 + auxiliary proficiency loss (weight 0.3)

4.1 Dataset Characteristics and Score Distribution

This section reports the characteristics of the ELLIPSE corpus obtained from the collected training and testing partitions. The distributions of scores, descriptive statistics, and relationships between the seven assessment dimensions provide the data context for the training and evaluation of SRAT for continuous multidimensional essay assessment.

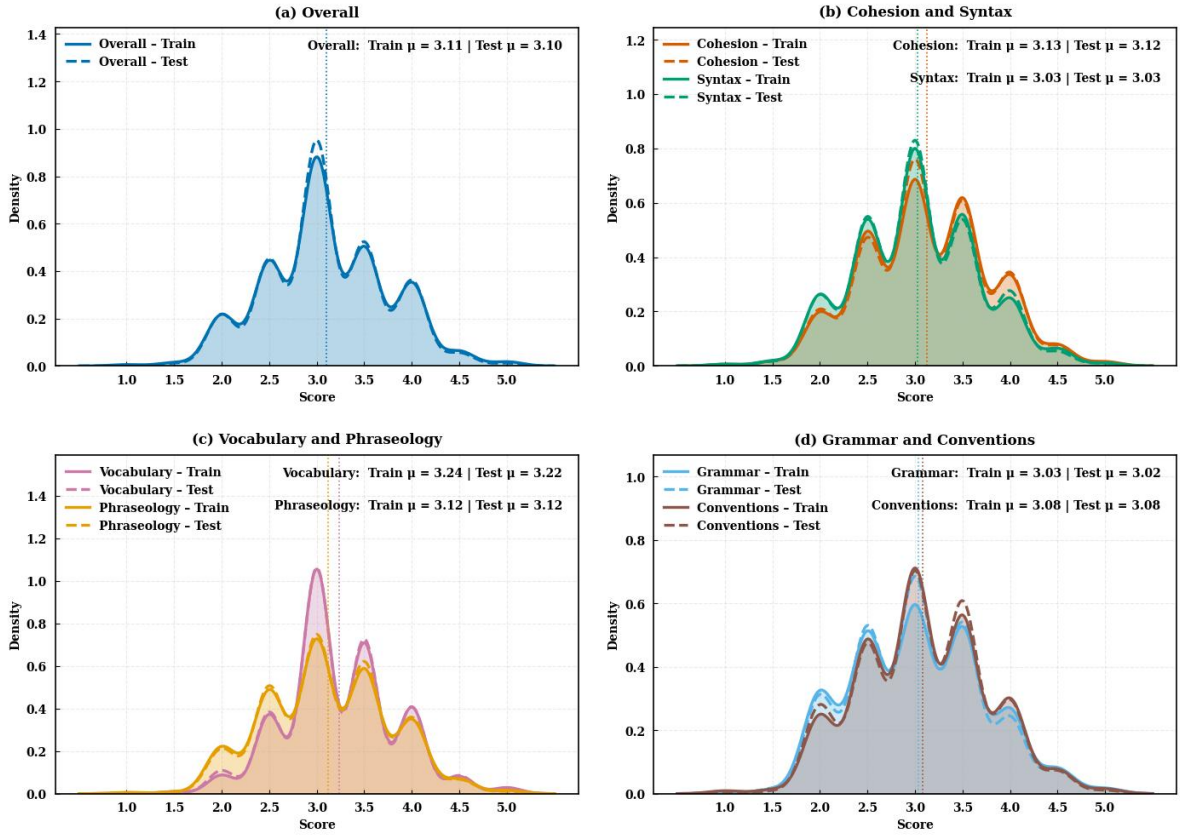

Figure 3. Score distributions of the seven ELLIPSE assessment dimensions

Figure 3 shows the distribution of essay scores by all seven assessment dimensions for the training and testing partitions. This analysis is key as the shape, spread and balance of the score range dictate the reliability with which a regression model can learn dimension specific patterns. Similar distribution on both partitions also suggest that the evaluation data do not cause distributional shift or sampling bias when evaluating the model. These evidences enable to train the dependable model.

Table 2. Descriptive statistics of ELLIPSE training and testing scores

Dimension	Train Mean	Train SD	Test Mean	Test SD	Min	Max
Overall	3.11	0.65	3.10	0.62	1.0	5.0
Cohesion	3.13	0.66	3.12	0.65	1.0	5.0
Syntax	3.03	0.64	3.03	0.64	1.0	5.0
Vocabulary	3.24	0.58	3.22	0.57	1.0	5.0
Phraseology	3.12	0.66	3.12	0.64	1.0	5.0
Grammar	3.03	0.70	3.02	0.67	1.0	5.0
Conventions	3.08	0.67	3.08	0.66	1.0	5.0

Table 2 shows summary statistics for each assessment dimension in the training and testing partitions. The measures are important because that help to identify the variability around the target variable as well as the range of scores for each target variable that indicates the relative difficulty of the target variable. That also show that both partitions are similar, which means that the choice of model is not influenced by the data and that the optimization of the continuous regression targets is stable, and that the prediction errors. These properties combine to give increased confidence in the evaluation protocol.

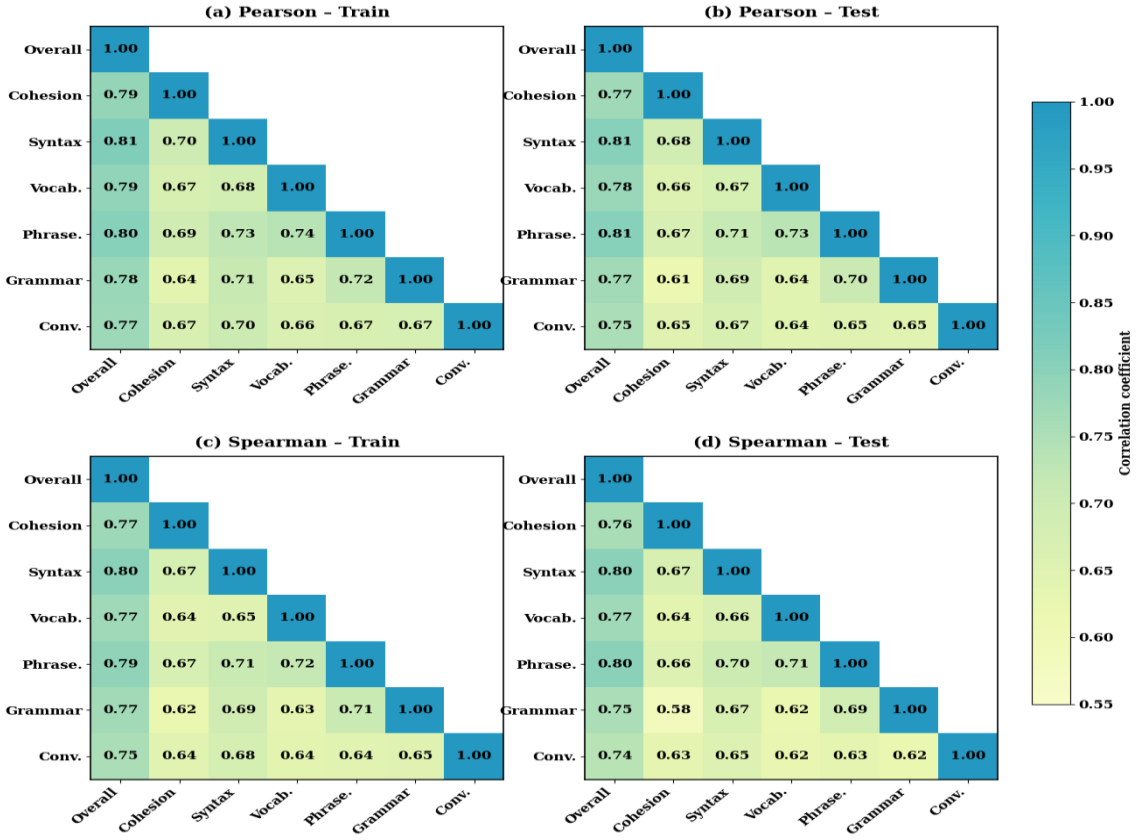

Figure 4. Correlation structure among the seven writing assessment dimensions

Figure 4 shows the linear and rank associated relationships between the seven writing assessment dimensions. This analysis is significant because the degree of association between the analytic skills is linked to the extent that that have the same information. The dimensions are strongly associated yet distinguishable, and therefore appropriate to be modeled together; consistent across partitions, and therefore appropriate to be generalized across partitions. This is the primary evidence to support the relational design of SRAT.

4.2 Preprocessing and Tokenization Results

This section presents the outcomes of input selection, attribute filtering, and independent tokenization of essay and writing-context texts. These results verify data validity and sequence-length suitability before the processed inputs enter the shared transformer encoder, confirming that essay content and context information are prepared consistently.

Table 3. Input filtering and tokenization statistics for model inputs

Parameter	Training	Testing
Records retained after filtering	3911.0	2571.0
Records removed (missing text/score)	0.0	0.0
Model input attributes retained	2.0	2.0
Target dimensions retained	7.0	7.0
Attributes excluded from inputs	17.0	17.0
Mean essay length (tokens)	486.7	479.2
Median essay length (tokens)	456.0	444.0
95th percentile essay length (tokens)	879.0	887.0
Maximum essay length (tokens)	1428.0	1504.0
Essays exceeding 512 tokens (%)	38.53	36.41
Mean context length (tokens)	5.3	5.3
Maximum context length (tokens)	9.0	9.0

Table 3 illustrates the effect of input selection, attribute filtering and subword tokenization on the essay and writing-context records. The significance of these statistics is that only valid linguistic inputs and target variables are kept and the sequences are of a length suitable for the capacity of the encoder. That also explain the maximum lengths of the inputs which were selected and add an estimation of how much of the information in

the essay is stored during tokenization. The counts are also consistent within the partitions, indicating that the processing prior to the count was the same for each partition.

4.3 Shared Contextual Encoding Results

This subsection considers the representations generated from the shared-weight DeBERTa-v3-base encoder for essays and writing context data. This study aims to assess whether the pooled features retain proficiency and task prompt information while maintaining compatibility within the same feature space.This underpins all subsequent modules.

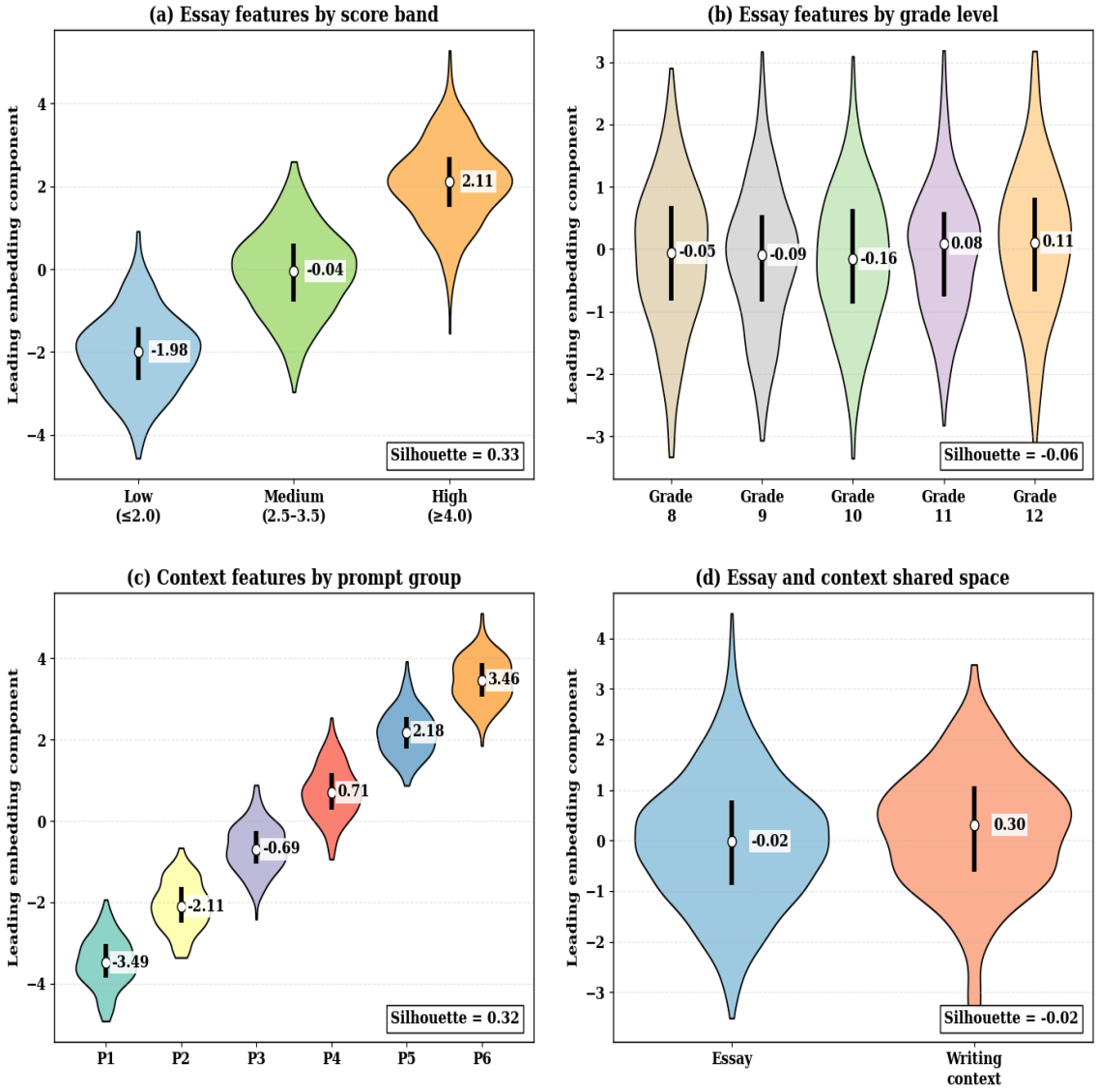

Figure 5. Embedding space visualization of shared encoder representations

Figure 5 illustrates the structure of the pooled essay and writing-context representations produced by the shared encoder. The purpose of this visualization is to see if the common feature space carries the information of proficiency, if unwanted demographic information does not cluster in the common feature space, and if the common feature space separates the writing contexts meaningfully. It gives evidence of how encoding with independent parameters is compatible and gives rise to compatible representations that are suitable for subsequent estimation of proficiency and context projection. The scoring down stream is dependable if these properties are exhibited.

4.4 Provisional Proficiency Estimation Results

In this section, the provisional estimates of proficiency produced by the pooling of essay representations are examined in relation to the holistic predictions that are eventually produced. The auxiliary estimate is shown to give the essay-based conditioning signal prior to attention refinement.

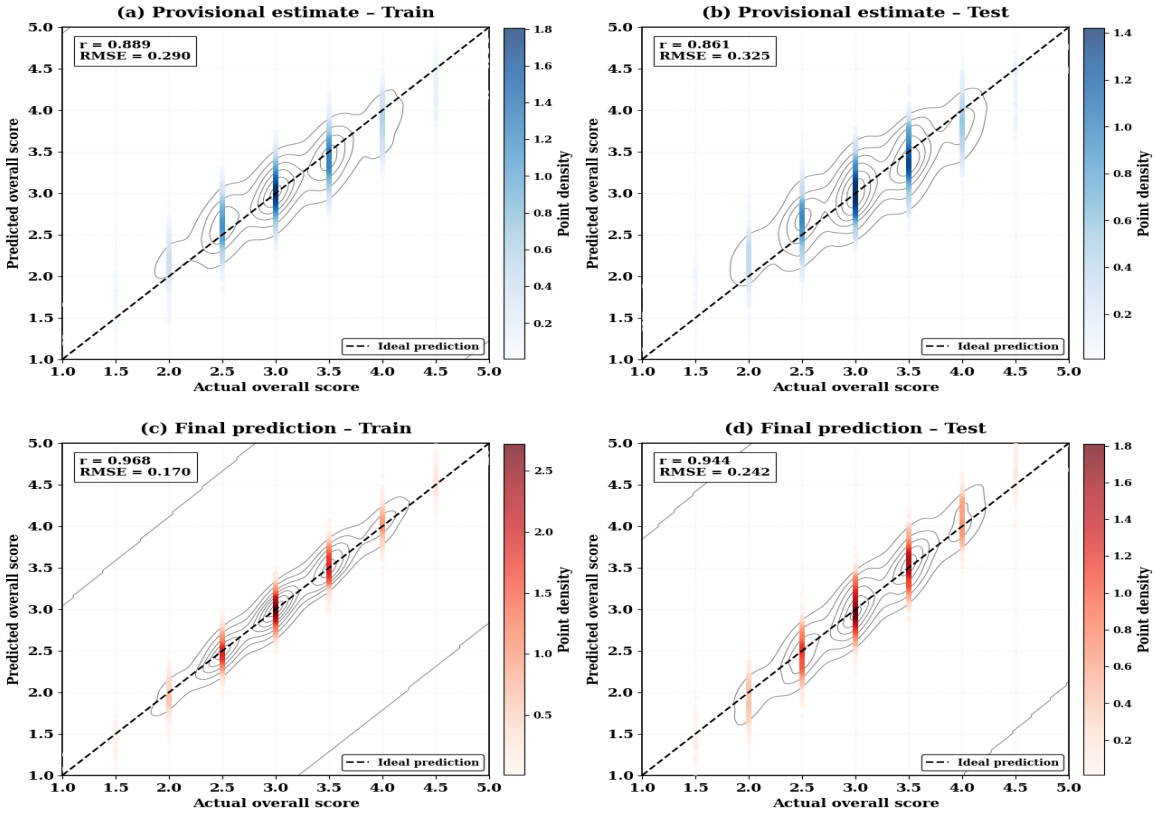

Figure 6. Provisional and final overall score density comparison

Figure 6 provides a graphical representation of the correlation of actual overall scores with the provisional proficiency estimates as well as the overall scores prediction. This is important because it provides proof that the conditioning signal derived from the auxiliary contains information regarding the essay, which can be enhanced further in subsequent modules. The density contours also show the prediction bias and spread of the scores. This proves validation to a strategy of predicting the scores in stages.

4.5 Proficiency-Conditioned Attention Results

This section examines the impact of the proficiency-dependent bias on token-level attention in the further multi-head component. The head-wise attentional activity will be examined against unconditioned attention to see whether proficiency conditioning brings attention to linguistically significant tokens pertaining to writing assessment. This validate the function of refinement.

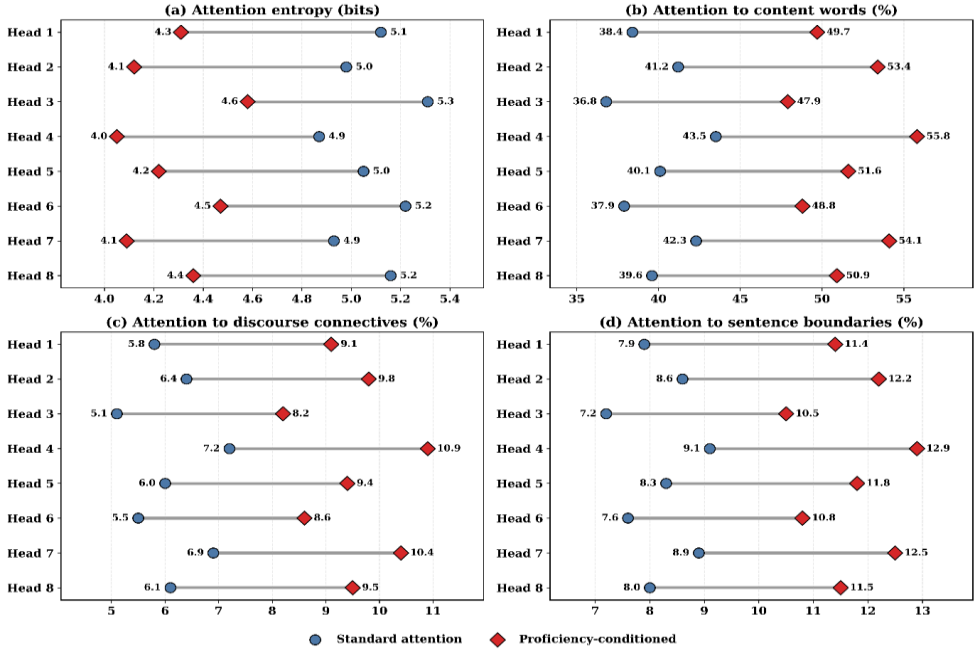

Figure 7. Head-wise attention behavior before and after proficiency conditioning

Figure 7 displays attention behavior of individual heads of the additional multi-head module following proficiency conditioning. This analysis is important since it validates the finding that the pairwise bias does not lead to a consistent effect, but rather pulls the reader's attention towards the informative content words, towards discourse connectives, and towards sentence boundaries. The case is that such selective refinement leads to the interaction with tokens being sensitive to the estimated ability to write. This evidence is in support of the proposed attention design.

4.6 Context Projection and Gated Integration Results

This subsection assesses the writing-context projection and the context-aware gated integration of refined essay features and projected context features.The learned gate values reveal how the model balances essay-derived and context-derived information at different proficiency levels within the shared representation.

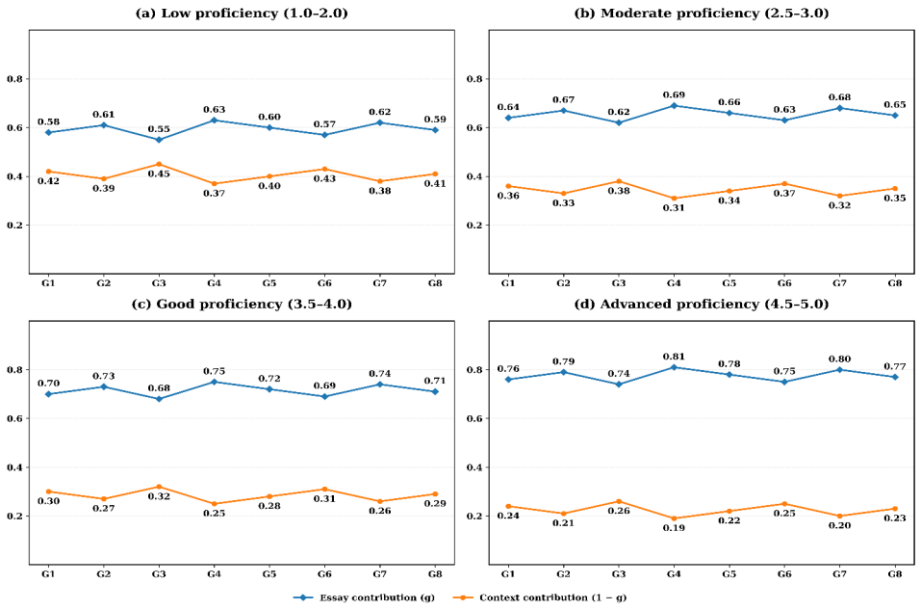

Figure 8. Gate-based essay and context contributions across proficiency levels

Figure 8 shows the relative contribution of refined essay features and projected writing-context features at proficiency levels. This is important as it reveals that the learned gate is able to adapt to feature selection and not fixed fusion weights. Knowing how much reliance students need to put on the essay and context information can help to confirm that integration module is acting sanely and adding complementary information. This behavior would lend credence to the adaptable fusion approach of SRAT and its strength in various writing qualities.

4.7 Skill-Relational Modeling Results

This section presents the relation weights learned among the six analytic writing dimensions. These weights indicate how information is shared across skills during relational refinement and whether the learned dependencies reflect linguistically meaningful associations between connected writing competencies. These weights guide residual feature refinement.

Table 4. Learned inter-skill relation weights among analytic dimensions

Target skill	Cohesion	Syntax	Vocabulary	Phraseology	Grammar	Conventions
Cohesion	0.26	0.19	0.15	0.17	0.13	0.10
Syntax	0.16	0.25	0.10	0.15	0.23	0.11
Vocabulary	0.14	0.09	0.27	0.28	0.12	0.10
Phraseology	0.15	0.13	0.26	0.22	0.13	0.11
Grammar	0.10	0.21	0.09	0.12	0.26	0.22
Conventions	0.09	0.12	0.08	0.12	0.27	0.32

Table 4 shows the row-normalized relation weights learned between the six analytic writing dimensions. This information is crucial as it is used to reveal the skills exchange information during the course of relational refinement, and to verify that each row is a valid distribution. If there are meaningful associations between grammar, conventions, vocabulary and phraseology, this suggests that the model reflects linguistically plausible dependencies, as opposed to arbitrary associations. The relations are valid to apply the residual relational refinement strategy, which was taken before the regression.

4.8 Training Behavior and Multidimensional Prediction Results

This section describes the optimization performance of the SRAT algorithm and the accuracy of its predictions in seven dimensions. While training curves demonstrate convergence, error and calibration analysis of results is used to evaluate the reliability and fairness of the regression results in different score bands and demographies.

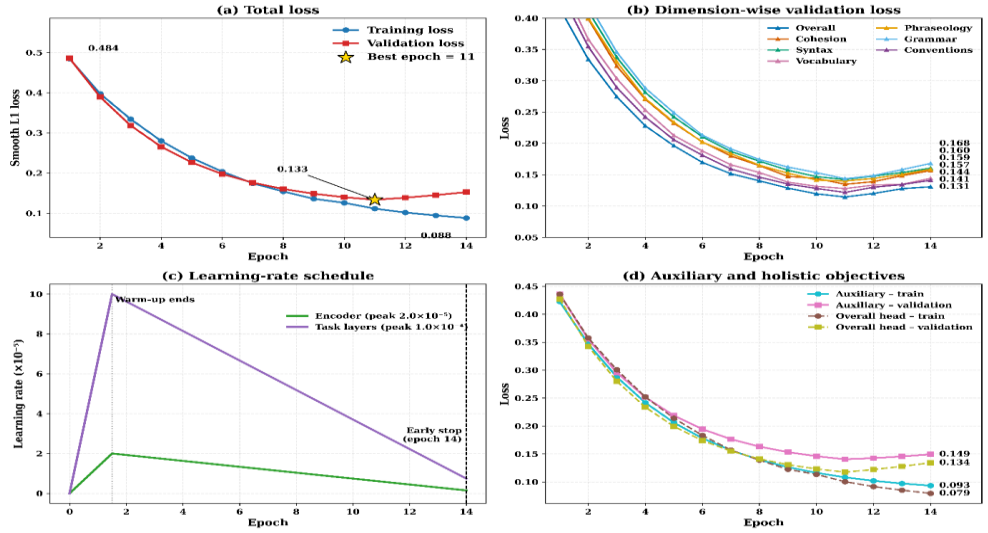

Figure 9. Training dynamics and optimization behavior of SRAT

Figure 9 shows the loss curves, learning-rate schedules and auxiliary objective behavior of SRAT training. Many of these curves are important to consider, such as making sure that they converge, showing signs of overfitting, and proving that the early stopping point was a good one. Moreover, the losses for each dimension have been smoothly reduced, indicating that the multi-objective regression heads are not trained separately and the training objective is not the dominant dimension loss. This is to guarantee the final model checkpoints and training configuration are reliable.

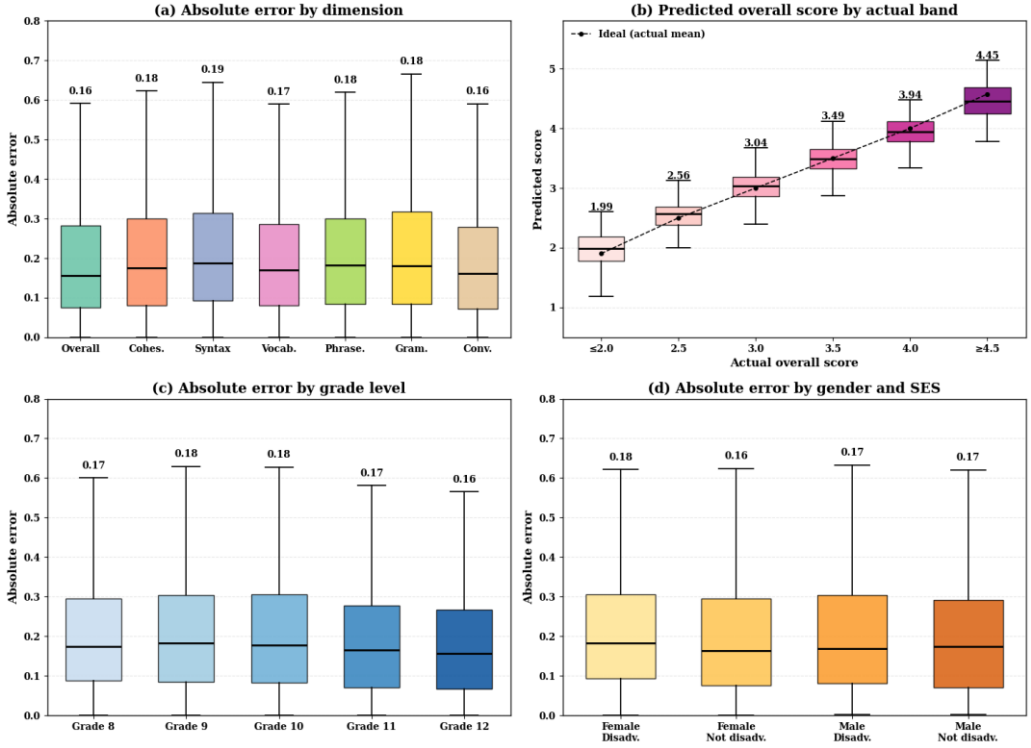

Figure 10. Prediction error and calibration analysis across groups

Figure 10 illustrates prediction errors and score calibration across dimensions, score levels, grades, and demographic groups. This analysis is valuable because overall performance may mask systematic bias at the high or low end or of a specific group of students. Reliable and equitable assessment are vital for implementing automated scoring in a variety of educational settings as shown by the balanced error distribution and aligned predicted scores. These results provide greater assurance of the utility of the framework.

4.9 Computational Efficiency Analysis

This section analyzes the parameters, computation, latency, and memory required by each architectural component of SRAT. The incremental analysis clarifies the practical cost of proficiency conditioning, gated integration, and relational modeling for deployment in educational assessment environments. Efficiency findings guide practical adoption.

Table 5. Computational cost of incremental SRAT architectural components

Configuration	Parameters (M)	GFLOPs per essay	Training time (min/epoch)	Latency (ms/essay)	Peak memory (GB)
Backbone only	183.8	90.6	5.9	14.8	14.2
+ Writing-context encoding	184.4	96.3	6.2	15.6	14.5
+ Proficiency-conditioned attention	187.5	98.6	7.3	17.4	21.8
+ Gated integration	188.7	98.8	7.4	17.7	22.0
+ Skill-relational modeling (full SRAT)	192.4	99.0	7.6	18.1	22.3

Table 5 presents the additional computation footprint added to the backbone encoder by each SRAT component. This information is relevant to practical deployment efforts in education as it relies on the number of parameters, computation, response time and memory requirements. Finally, quantifying the overhead of proficiency conditioning, gating and relational modeling illustrates if the proposed gains in architecture are realized at a cost that is acceptable and justifiable. The results inform on-going implementation in resource-constrained settings.

4.10 Ablation Study

This section examines the contribution of individual SRAT components by removing or replacing them while keeping all other settings unchanged. The analysis identifies which modules contribute most to scoring performance and confirms the complementary roles of context, proficiency conditioning, gating, and relational modeling.

Table 6. Ablation results for individual SRAT architectural components

Variant	QWK	RMSE	MAE	PCC
Backbone only	0.852	0.336	0.268	0.865
w/o Writing context	0.878	0.307	0.245	0.889
w/o Proficiency-conditioned attention	0.889	0.292	0.233	0.899
w/o Gated integration	0.896	0.282	0.225	0.906
w/o Skill-relational modeling	0.901	0.274	0.219	0.911
w/o Auxiliary proficiency supervision	0.905	0.267	0.213	0.915
Full SRAT	0.912	0.258	0.206	0.921

Table 6 shows the ablation study results of the proposed SRAT framework, indicating the impact of the individual components on the overall performance. Results indicate that the removal of any component would cause a decrease in the value of QWK and PCC and increase RMSE and MAE. The integrated architecture delivers the best performance, with a QWK of 0.912, RMSE of 0.258, MAE of 0.206, and PCC of 0.921.

4.11 Performance Comparison

To evaluate the effectiveness of the proposed SRAT framework, its performance is compared with recently published automated essay-scoring approaches. Four commonly used evaluation metrics, including QWK, RMSE, MAE, and PCC, are considered to examine scoring agreement, prediction errors, and correlation performance across the selected methods.

Table 7. Comparative Performance of Automated Essay-Scoring Methods

Study / Model	QWK	RMSE	MAE	PCC
IRT-AESF (Prompt 4)[23]	0.804	0.583	0.321	0.806
Ontology + T2NS + Spatial Autocorrelation[24]	0.870	0.630	0.470	0.910
Error-Supervised Synthetic Learner (Authentic data)[25]	0.701	0.449	0.361	--
Proposed SRAT	0.912	0.258	0.206	0.921

Table 7 compares the proposed SRAT framework with three existing essay-scoring approaches. The results show that the highest QWK (0.912) and PCC (0.921) as well as the lowest RMSE (0.258) and MAE (0.206) are

obtained by SRAT. The results reported here suggest that the accuracy in the scoring agreement, prediction accuracy and error reduction are improved, compared to the results reported in existing approaches.

4.12 Discussion

The ablation results indicate that each of the SRAT components makes a meaningful contribution to scoring quality. Removing writing-context encoding resulted in the most decrease, and sequentially, proficiency-conditioned attention, gated integration, skill-relational modeling, and auxiliary supervision were shown to be complementary. The full model had increased agreement and correlation and reduced prediction error compared to the backbone. The relation weights learned were associated with vocabulary and phraseology with grammar, and the conventions with grammar, illustrating plausible linguistic relationships between writing skills. Analysis of essay and topic information control in the gate behavior across proficiency levels revealed that the model adapted essay and topic information rather than fused them, which was predetermined. A modest increase in parameters, computation, and latency has resulted in these benefits, which make them suitable for practical applications in educational assessment. These results support the integrated design as compared to individual modules.

5. Conclusion and future works

This study presented SRAT, a context-aware framework for multidimensional English essay assessment that combines shared DeBERTa-v3-base encoding, proficiency-conditioned attention, adaptive gated integration, and skill-relational modeling. The complete model showed the best agreement, correlation and error rates for the ELLIPSE corpus, and ablation results showed that each component helped improve these scores. The learned relation weights also showed linguistically plausible relations between grammar, conventions, vocabulary, and phraseology without incurring too great a computational burden. The results show that knowledge of the skill is more informative and coherent than the backbone encoder alone when combined with knowledge of the proficiency and topic context.

Future research will build on the SRAT framework by adding full prompt instructions to further develop prompt-sensitive assessment and investigate effective encoding of long essays. The framework will also be tested for ongoing adaptation of newly introduced prompts and for generalization across prompts and genres. Other directions include the creation of lightweight variants of the attention mechanism for deployment in resource-constrained environments, fairness and calibration analysis across different groups of learners and educational settings, and the validation against human raters for facilitating reliable use in classrooms.

References

[1] N. M. Bui and J. S. Barrot, “ChatGPT as an automated essay scoring tool in the writing classrooms: how it compares with human scoring,” *Educ Inf Technol*, vol. 30, no. 2, pp. 2041–2058, Feb. 2025, doi: 10.1007/s10639-024-12891-w.

[2] W. Villegas-Ch, J. Govea, R. Albuja, D. Buenaño-Fernández, and A. Mera-Navarrete, “Optimizing Automated Educational Systems: Adaptive Syllabus Generation and Precise Assessment Using Artificial Intelligence,” *IEEE Access*, vol. 12, pp. 173390–173409, 2024, doi: 10.1109/ACCESS.2024.3503532.

[3] Z. Du and K. Hashimoto, “AcademiCraft: Transforming Writing Assistance for English for Academic Purposes with Multi-Agent System Innovations,” *Information*, vol. 16, no. 4, p. 254, Mar. 2025, doi: 10.3390/info16040254.

[4] W. Xu, R. Mahmud, and W. Lam Hoo, “A Systematic Literature Review: Are Automated Essay Scoring Systems Competent in Real-Life Education Scenarios?,” *IEEE Access*, vol. 12, pp. 77639–77657, 2024, doi: 10.1109/ACCESS.2024.3399163.

[5] H. Misgna, B.-W. On, I. Lee, and G. S. Choi, “A survey on deep learning-based automated essay scoring and feedback generation,” *Artif Intell Rev*, vol. 58, no. 2, p. 36, Dec. 2024, doi: 10.1007/s10462-024-11017-5.

[6] S. Tzimiris, S. Nikiforos, M. N. Nikiforos, D. Mouratidis, and K. L. Kermanidis, “A Comparative Evaluation of Transformer-Based Language Models for Topic-Based Sentiment Analysis,” *Electronics*, vol. 14, no. 15, p. 2957, Jul. 2025, doi: 10.3390/electronics14152957.

[7] Z. Zhao, Y. Wang, J. Li, and S. Che Lah, “A Calibrated Multi-LLM Consensus Framework for Reducing Model-Specific Scoring Asymmetry in EFL Writing Assessment,” *Symmetry*, vol. 18, no. 10, p. 1714, Oct. 2026, doi: 10.3390/sym18101714.

[8] M. Cho, J. Huang, and O. Kwon, “Dual-scale BERT using multi-trait representations for holistic and trait-specific essay grading,” *ETRI Journal*, vol. 46, no. 1, pp. 82–95, Feb. 2024, doi: 10.4218/etrij.2023-0324.

[9] F. Lin, “CNN–transformer-based trait-aware scoring and student-aware feedback generation for English writing,” *Discov Artif Intell*, vol. 6, no. 1, p. 405, Mar. 2026, doi: 10.1007/s44163-026-01132-3.

[10] A. Li, “Research on English text scoring technology based on deep learning in English teaching,” *Discov Artif Intell*, vol. 6, no. 1, p. 90, Jan. 2026, doi: 10.1007/s44163-025-00790-z.

[11] C. Impey, M. Wenger, N. Garuda, S. Golchin, and S. Stamer, “Using Large Language Models for Automated Grading of Student Writing about Science,” *International Journal of Artificial Intelligence in Education*, vol. 35, no. 4, pp. 1825–1859, Dec. 2025, doi: 10.1007/s40593-024-00453-7.

[12] J. Choi, T. Tate, and M. Warschauer, “Anchor is the key: Toward accessible automated essay scoring with large language model through prompting,” *Assessing Writing*, vol. 69, p. 101053, Jul. 2026, doi: 10.1016/j.asw.2026.101053.
[13] H.-N. Wu, M.-N. Chu, and J.-L. Hsu, “Comparing GPT and human raters in essay assessment: Variability, bias, and the potential of LLM-based scoring,” *Computers and Education Open*, vol. 10, p. 100341, Jun. 2026, doi: 10.1016/j.caeo.2026.100341.
[14] J. Sun, W. Peng, T. Song, and J. Song, “UniMTES: a unified framework with trait-description-aware for multi-trait essay scoring,” *Complex Intell. Syst.*, vol. 12, no. 3, p. 98, Mar. 2026, doi: 10.1007/s40747-025-02224-w.
[15] D. Han, D. Roh, E. Han, H. Song, and M. Y. Yi, “Fine-grained multi-prompt essay scoring with multi-level disentanglement,” *Data Min Knowl Disc*, vol. 39, no. 5, p. 67, Sep. 2025, doi: 10.1007/s10618-025-01126-2.
[16] Y. Ren, W. Fan, and J. Wang, “Intelligent text analysis for effective evaluation of english Language teaching based on deep learning,” *Sci Rep*, vol. 15, no. 1, p. 28949, Aug. 2025, doi: 10.1038/s41598-025-14320-5.
[17] T. Song, J. Sun, and W. Peng, “Enhancing Cross-Prompt Essay Trait Scoring via External Knowledge Similarity Transfer,” *Symmetry*, vol. 17, no. 5, p. 739, May 2025, doi: 10.3390/sym17050739.
[18] T. Miao and D. Xu, “KWM-B: Key-Information Weighting Methods at Multiple Scale for Automated Essay Scoring with BERT,” *Electronics*, vol. 14, no. 1, p. 155, Jan. 2025, doi: 10.3390/electronics14010155.
[19] J. F. Lohmann *et al.*, “Neural Networks or Linguistic Features? - Comparing Different Machine-Learning Approaches for Automated Assessment of Text Quality Traits Among L1- and L2-Learners’ Argumentative Essays,” *International Journal of Artificial Intelligence in Education*, vol. 35, no. 3, pp. 1178–1217, Sep. 2025, doi: 10.1007/s40593-024-00426-w.
[20] X. Tang, H. Chen, D. Lin, and K. Li, “Harnessing LLMs for multi-dimensional writing assessment: Reliability and alignment with human judgments,” *Heliyon*, vol. 10, no. 14, p. e34262, Jul. 2024, doi: 10.1016/j.heliyon.2024.e34262.
[21] A. Pack, A. Barrett, and J. Escalante, “Large language models and automated essay scoring of English language learner writing: Insights into validity and reliability,” *Computers and Education: Artificial Intelligence*, vol. 6, p. 100234, Jun. 2024, doi: 10.1016/j.caeai.2024.100234.
[22] S. Crossley, “scrosseye/ELLIPSE-Corpus: the English Language Learner Insight, Proficiency and Skills Evaluation (ELLIPSE) Corpus.” Accessed: Nov. 06, 2026. [Online]. Available: https://github.com/scrosseye/ELLIPSE-Corpus
[23] W. Xia, J. Wu, J. Yu, and C. Zheng, “The Best of Two Worlds: IRT-Enhanced Automated Essay Interpretable Scoring,” *Behavioral Sciences*, vol. 16, no. 4, p. 542, Apr. 2026, doi: 10.3390/bs16040542.
[24] S. M. Darwish, N. A. Bagi, and N. A. El-Shoafy, “Intelligent automated essay scoring under uncertainty using type 2 neutrosophic ontologies,” *Sci Rep*, vol. 16, no. 1, p. 17186, Jun. 2026, doi: 10.1038/s41598-026-54596-9.
[25] D. A. Nguyen, “Error-Supervised Synthetic Learner Writing for Automated Essay Scoring,” Sep. 22, 2026, *arXiv*. doi: 10.48550/ARXIV.2609.23573.