SCIENTIFIC CULTURE
www.sci-cult.com

SCIENTIFIC CULTURE, Vol. 12, No. 4, (2026), pp. 10413-10428
Open Access. Online & Print

DOI: 10.5281/zenodo.124261008

PROFICIENCY-AWARE EXPLAINABLE TRANSFORMER FRAMEWORK FOR FAIR AND RUBRIC-ALIGNED AUTOMATED ESL WRITING EVALUATION

Gulnaz Fatma[1]*

[1]Department of English Al Diar University College Jazan University Jazan, Saudi Arabia

Received: **29/12/2025**
Accepted: **20/04/2026**

Corresponding Author: Gulnaz Fatma (gulnaz.fatima15@gmail.com)

ABSTRACT

Automated evaluation of English as a Second Language (ESL) writing is critical for delivering personalized learning at scale, yet traditional assessment methods remain limited by subjectivity, high resource demands, and low interpretability. Existing automated essay scoring (AES) systems often overlook rubric-specific dimensions, learner proficiency levels, and fairness across diverse student subgroups, resulting in inconsistent and potentially biased evaluations. To address these challenges, this study introduces the Proficiency-Aware Explainable Writing Evaluation Framework (PA-EWEF), which integrates transformer-based text encoding, proficiency-calibrated discourse modeling, dynamic rubric weight learning, explainability layers, and fairness-aware optimization. Implemented in Python and PyTorch, the framework performs tokenization, attention-guided representation learning, and rubric-aligned score prediction across content, organization, and language dimensions, while generating saliency maps and attention visualizations for interpretable feedback. Experimental ***results*** demonstrate that PA-EWEF outperforms conventional machine learning models—including Random Forest, Support Vector Regression, AdaBoost, and XGBoost—achieving an MAE of 0.147 and RMSE of 0.173, surpassing previously reported ranges (MAE 0.152–0.787; RMSE 0.168–1.338). Rubric-level and fairness analyses confirm consistent performance across proficiency subgroups and highlight the framework's ability to provide actionable pedagogical insights. These findings validate PA-EWEF as an accurate, interpretable, and equitable solution for automated ESL writing assessment, underscoring the value of combining explainable AI, adaptive rubric modeling, and fairness optimization to enhance reliability, transparency, and instructional utility.

Keywords— Automated Essay Scoring (AES), ESL Writing Assessment, Explainable AI (XAI), Transformer Models, Proficiency-Aware Evaluation, Rubric-Aligned Scoring, Fairness-Aware Optimization, Attention Visualization

1 INTRODUCTION

Automated assessment of English as a Second Language writing has gained greater significance in mass education, online learning systems and formative assessment systems [1], [2]. The traditional human grading is pedagogically rich but has its limitation due to subjectivity, high cost and scalability issues [3]. The first automated essay scoring systems were based on manually engineered linguistic representations and rule based scoring, with little scalability between proficiency levels of learners [4], [5]. Recent discoveries in transformer-based language models have greatly enhanced the contextual interpretation of writing by learners, as they are syntactically, semantically, and discourse-based comprehensive. These models have ability to perform fine-grained scoring and are more accurate in prediction than the traditional neural methods. These systems are generally however opaque black boxes, which supports less interpretability and pedagogical confidence and utilize fixed scoring rubrics [6]. In addition to this, homogenization of different ESL learners tends to increase prejudice, especially between proficiency levels and learning phases [7]. Although explainable artificial intelligence methods have been proposed to respond to transparency, they are not often consistent with analytic rubrics applied in the actual educational context. Consequently, the existing systems of automated writing assessment are unable to reconcile accuracy, fairness, adaptability and instructional relevance in one system.

Recent studies have been conducted on rubric based scoring, attention visualization and fairness conscious learning to enhance the accuracy of automated ESL writing assessment [8]. Multidimensional feedback can be achieved using analytic rubric modeling, whereas partial insight into model decision-making can be acquired using attention mechanisms [9]. However, the vast majority of the current research approaches rubric learning, explainability, and fairness as the separate goals instead of the aspects that rely on one another [10]. The use of fixed rubric weights does not indicate the changing levels of learner proficiency, and explainability practices tend to be out of touch with pedagogically significant sets of criteria [11]. Moreover, transformer encoders usually produce the representations without taking explicit consideration of the latent proficiency-driven discourse patterns, which makes the linguistic features and scoring justification misaligned [12]. These disadvantages limit the use of automated assessment in real classroom situations where flexibility and openness are needed. To overcome these issues, this study proposes a single evaluation system, which combines proficiency-sensitive representation learning, adaptive rubric modeling, and fairness-conscious optimization. The suggested solution makes use of transformer-based encoding with discourse calibration and explainability mechanisms. This study presents a new proficiency-aware writing assessment framework, Proficiency Aware Explainable Writing Evaluation Framework, a hybrid between dynamic rubric learning and a new proficiency-aware discourse calibration layer that allows flexibly, equally, and understandably assessing ESL writing.

A. Problem Statement

Current automated ESL writing assessment software suffer a number of severe limitations which limit their application in education. The formulations of static rubrics do not follow a heterogeneous learner level of proficiency, and the scoring behavior is not consistent across the stages of development [13]. The scoring model, though correct, is not explicitly aligned to the analytic rubrics and has low levels of pedagogical transparency, which is a drawback of transformer-based scoring models [14]. Explainability methods are mostly shallow, with token-level significance, but no real connection with teaching standards [15]. The methods that are fairness-conscious mitigate bias after the fact instead of incorporating equity principles in the representation learning process. Besides, the vast majority of research ignores discourse progression patterns that characterize ESL writing and lead to the disproportionate focus on surface-based linguistic characteristics. All these inefficiencies lead to a lack of trust, interpretability, and instructional usefulness. There is a distinct knowledge gap of creating an integrative framework jointly modelling proficiency-sensitive discourse representations, learning dynamically the importance of the rubric, balancing fairness in subgroups of learners and providing rubric-aligned explanations in one transformer-based system.

B. Research Motivation

The reason behind this study is that there is a necessity to develop automated ESL writing evaluation systems that can be in accordance with the actual classroom assessment procedures and at the same time be transparent and fair. The growing use of AI-based assessment will require the models to be adaptable to learner proficiency, offer interpretable feedback and reduce bias. The communication between high-level language modelling and pedagogically based rubric assessment is still needed in order to have reliable and scalable educational implementation.

C. Significance of the Study

The study provides a contribution to the field of educational artificial intelligence through the creation of a framework that incorporates adaptability, explainability, and fairness in ESL writing assessment. Discourse modeling that is sensitive to proficiency increases the reliability of assessment at different stages of learning and dynamic learning in rubrics promotes the achievement of instructional goals. The approach proposed provides fair-minded scoring and understandable feedback, and it has practical merits to the teachers, institutions, and smart tutoring systems that need reliable automated assessment tools.

D. Key Contributions

- Introduces a proficiency-aware discourse calibration layer that modulates transformer representations according to latent ESL proficiency signals.
- Develops a dynamic rubric learning mechanism that adaptively assigns analytic rubric importance during model training.
- Integrates rubric-aligned explainability to generate interpretable feedback grounded in pedagogically meaningful criteria.
- Incorporates fairness-aware optimization to reduce performance disparities across learner proficiency subgroups.
- Validates the proposed framework on large-scale rubric-annotated ESL data, demonstrating improvements in accuracy, equity, and interpretability.

E. Rest of the sections

The remainder of the study is organized as follows. Section II provide an extensive review of the literature that discusses the existing models and their limitations. Section III, elaborate on the proposed approach. Section IV present the experimental results and interpretation. The conclusion and future research directions, limitations, and recommendations are discussed in Section V.

2 LITERATURE REVIEW

Previous research by Almusharra and Alotaibi [16] offers an empirical basis to the comprehension of the correspondence between automated essay scoring (AES) and human rating in EFL settings. In their analysis, they have emphasized that the commercial AES tools have the potential to substantially decrease workload on grading and detect linguistic errors in a systematic way. Simultaneously, the results of their studies reveal a systematic leniency bias in scoring as done by people, and they show that AES systems are more likely to pick up superficial errors. This text highlights a significant drawback of the initial AES studies: reliability is able to be measured in a statistical manner, but the question of pedagogical transparency and reasoning at the rubric level are not much addressed, which limits the interpretability of the classroom.

Developing the semantic constraints of traditional AES, Darwish et al. [17] redirect the scope of interest towards surface accuracy to a meaning-based evaluation. Their combination of the latent semantic analysis and neutrosophic ontology proves that uncertainty and vagueness on language explicitly can be modeled. The analysis reveals that there were evident improvements in semantic scoring accuracy and richness of feedback over the baseline systems. Nonetheless, the use of handcrafted ontologies creates scalability issues and indicates a larger trade-off between semantic control and open-domain writing adaptability in AES research studies.

Xuan [18] provides an opposite course of action by focusing on scalability with deep neural architectures. This work shows that contextual and sequential dependencies can be collectively used to enhance grammar and coherence prediction by using transformers together with LSTMs. The discussion makes hybrid neural models' good candidates to large-scale application. However, the lack of analytic rubric design and clear fairness modeling indicates a repetitive drawback of purely data-driven methodology high performance does not imply clear or pedagogical based assessment.

Making less emphasis on scoring and more emphasis on the learner support, Zheng and Zhang [19] examine the writing assistance with the help of the multidimensional, real-time feedback system. Their work demonstrates that the use of transformer-based models can provide a variety of adaptive feedback at various levels of linguistics with minimal latency. It is possible to note that the analysis demonstrates the objective change in the quality of writing, but points to the lack of conceptualization of the relationships between formative feedback systems and summative assessment. Such systems are not easy to implement in standardized assessment environments, unless they have rubric-based scoring or proficiency modeling.

The works of language modeling by Škorić et al. [20] provide knowledge on the morphology-rich languages and evidence that the composite generative models are superior to the single models in text classification. Their results are in line with the efficiency of perplexity-based stacked classifiers at

language quality discrimination. Nevertheless, the absence of educational purpose, modeling of learners and of evaluation based on rubrics limits the relevance of such methods to ESL writing evaluation which strengthens the gap between the general NLP developments and the assessment requirements of education.

In the context of the assessment theory, Peeters and Gonyeau [21] discuss the efficacy of the analytic and the mixed-method rubrics. Their results contradict the belief that analytic rubrics are superior at all times, demonstrating that mixed rubrics could be equally reliable with significantly less effort. In spite of being performed by hand, this study has analytical importance in that it points to the design of rubrics as a critical variable, which is either over-simplified or ignored by most automated systems. Jiang and Zhang [22] are more explicit in their explanation of explainability in assessment, studying the interpretation of quality in explainable machine learning. Their framework indicates that clear feature attribution has the ability to endorse diagnostic feedback and construct validity. Nevertheless, reliance on the hand-coded qualities and organized data exposes the constraints to working with the more complex discourse and contextual language phenomena that the focus of writing assessment.

Mastour et al. [23] also generalize explainable analytics to high-stakes educational assessment and show that ensemble models with SHAP can predict with near-perfect accuracy. Their discussion demonstrates the importance of interpretability to institutional decision-making. However, the structured and domain specificity of the data limits transferability such that explainability is not enough but rather demands models that can process both unstructured writing data.

The limitations that were identified in the literature are addressed systematically in this study to a single, rubric-conscious, and explainable transformer-based framework of ESL writing evaluation. The suggested technique, contrasting with the previous designs that involve fixed criteria, handcrafted features, or domain-driven designs, is the introduction of the dynamic rubric learning that can adjust scoring weights depending on the proficiency of the learner and the discourse context. Deep contextual representations Fetching out of transformer encoders surmount surface-level and ontology-specific constraints, whereas integrated explainability layer connects attention attribution to the dimensions of the pedagogical rubric to achieve transparency. Moreover, a fairness-conscious optimization will reduce the subgroup bias that previous studies have ignored, and make fair evaluation of different learner groups possible. The combination of adaptivity, interpretability, fairness, and discourse-level modeling in the proposed framework is a means to go beyond such individual improvements and towards a unified, scalable, and pedagogically-based writing assessment paradigm.

TABLE I. summary of existing literature

Reference	Method	Advantages	Limitations
Almusharra and Alotaibi [16]	Commercial AES evaluated via error analysis and statistical tests	Reduces grading effort; reveals human–AES scoring differences	Proprietary tool; no rubric modeling, explainability, or proficiency awareness
Darwish et al. [17]	Latent Semantic Analysis with neutrosophic ontology	Improves semantic and syntactic accuracy; manages uncertainty	Handcrafted ontology; poor scalability; no adaptive rubrics
Xuan [18]	Hybrid Transformer–LSTM scoring and feedback model	Strong grammar and coherence prediction; scalable architecture	Static criteria; lacks rubric alignment, fairness, and transparency
Zheng and Zhang [19]	Transformer-based multidimensional feedback system	Real-time, multi-level feedback; low latency	No formal scoring; missing rubrics, fairness, and proficiency modeling
Škorić et al. [20]	Composite generative models with perplexity-based classifiers	Effective classification in morphology-rich languages	Not educational; no rubrics, learner modeling, or explainability
Peeters and Gonyeau [21]	Analytic vs. mixed rubric comparison using Rasch analysis	Efficient and reliable rubric evaluation	Manual only; no automation or adaptive scoring
Mastour et al. [23]	Explainable ensemble ML with SHAP	High accuracy; strong interpretability	Domain-specific; structured data only; not language-based
Jiang and Zhang [22]	Explainable ML with feature engineering and SHAP	Transparent predictions; diagnostic feedback	Handcrafted features; no deep discourse or dynamic rubrics

Table I demonstrates a comparative overview of the representative studies in automated assessment and rubric-based evaluation with emphasis on their methodological basis, main benefits, and limitations they have. It compares in an organized cycle traditional AES system, neural and hybrid models, ontology-directed and explainable frameworks of learning. The comparisons indicate that there is

continued lack of proficiency-sensitive model, adaptation of rubric dynamism, integration of fairness, and adaptive explainability, thus inspiring the incentive to consolidate and dynamic ESL writing evaluation system.

3 PROPOSED METHOD FOR PROFICIENCY AWARE EXPLAINABLE WRITING EVALUATION FRAMEWORK

The framework set out is an end-to-end, adaptive, automated system of the evaluation of ESL writing that simultaneously considers scoring expertise, interpretability, and fairness. The system commences with strict data preprocessing to maintain timely relevancy, consistency of annotations as well as distributional balance. Essays are then coded with a text encoder based on transformer which encodes token-level linguistic information, sentence-level semantic and discourse-level coherence. These representations are then further elaborated using a proficiency aware discourse calibration mechanism that allows responsiveness to developmental stages of the learners. Dynamic rubric learning, as an alternative to fixed scoring rules, estimates content, organizational and linguistic rubric significance in instances. An explainability layer converts internal attention dynamics to the rubric-based feedback, whereas fairness-sensitive optimization reduces the gap between the performance of subgroups. A combination of these elements constitutes one continuous evaluation pipeline, which resembles pedagogical reasoning without losing its computational power.

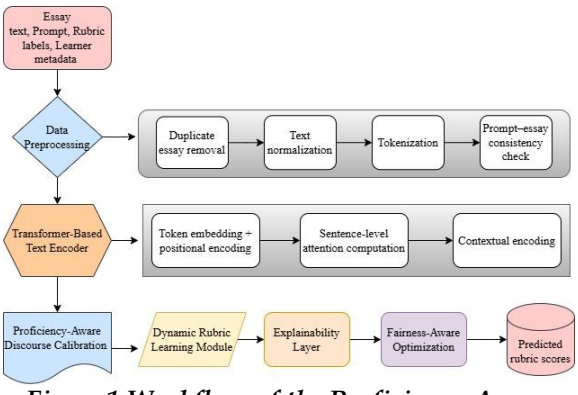

Figure 1 Workflow of the Proficiency-Aware Explainable Writing Evaluation Framework

Fig. 1 illustrates the workflow. It starts with preprocessing of the essays, with cleaning, normalization and tokenization, and encoding of the token, sentence and discourse features with transformers. Proficiency-aware calibration adjusts discourse representation, dynamic rubric learning predicts adaptive scores, explainability generates rubric-aligned feedback, and fairness-aware optimization minimizes subgroup disparities, producing accurate, interpretable, and equitable ESL assessment outputs.

F. Data Collection

The DREsS data set is designed in such a way that data is transparent, methodological and reproducible in rubric-based ESL writing assessment [24]. It consolidates three complementary subsets: DREsS_New, DREsS_Std., and DREsS_CASE which gives an overall number of about 49,000 essay samples. DREsS New is a collection of genuine in-classroom ESL argumentative essays on answers to critically crafted prompts and graded by analytic rubrics on a standardized scoring scale. DREsS_Std. standardizes several existing AES datasets on common rubric definitions to lessen the effect of noise across datasets and improve comparability. DRES_CASE adds even more data by using the controlled synthetic essay generation, increasing the variety of the data without denying the pedagogical validity. The essays are rated on the content dimension, organization dimension as well as language domain, allowing modeling finer-grains, rubric-aligned.

G. Data Preprocessing

Preprocessing of data makes it consistent, valid and robust before the corpus of ESL essays is trained on the model. Essays are processed with duplicates and non-completes being filtered out and then text normalizing is done which does not remove ESL-based linguistic cues. Essay alignment is taken timely to ensure relevancy of tasks. Transformer compatible tokenization with attention masking is used and the datasets are stratified into training, validation and testing partitions depending on the rubric scores and proficiency levels to minimize bias.

1) Data Cleaning and Validation: All essay samples are systematically filtered to eliminate duplicates, incomplete submissions and corrupted text entries that occur due to collect or format errors. The rubric-level scores are checked to be true to the specified scoring scale, and the discrepancies between the rubric components and the final scores are either corrected or omitted. The latter ensures the reliability of annotation and eliminates the risk of noisy labels affecting the learning of a model.

2) Text Normalization: Control of lowercasing, punctuation standards, and sentence division detectors are used to normalise essays so that the

same textual representation is achieved. Such aspects as ESL-specific grammatical structures, spelling differences, and the mistakes made by the learners are not eliminated (in contrast to aggressive normalization), because these aspects are pedagogically important in terms of rubric-based assessment and modeling that is sensitive to proficiency. One key data preprocessing method is the process of normalizing scores, which is used to allow the learning process to be equally learnable across rubric dimensions, which have variable value ranges. Rubric scores in this study are normalized with the help of min max scaling is expressed in eqn (1).

$$s_i^{\text{norm}} = \frac{s_i - s_{\min}}{s_{\max} - s_{\min}} \quad (1)$$

Where, s_i id denotes the original rubric score of an essay and $s_{\min}$, $s_{\max}$ denotes the minimum and maximum possible rubric score, respectively. Such normalization stabilizes optimization, does not allow dominance of one of the dimensions of a rubric, and provides equal contribution to the dynamic learning of the rubric.

3) Prompt–Essay Consistency Check: All essays are checked against their prompts to ensure that they are topical and that they comply with the task. Semantic similarity checks are used to detect off-topic or poorly aligned responses and are removed to eliminate the confounding effects associated with content and organization scoring dimensions.

4) Tokenization and Encoding: Transformer-adaptable tokenizers convert validated essays into subword-representations. Fixed sequence length and attention masks are used to process the variable size of essays with long-range dependencies that are important to discourse-level analysis.

5) Stratified Dataset Splitting: Stratification is used to divide the dataset into training, validation, and testing to separate the dataset based on the rubric scores and the levels of proficiency. The strategy ensures consistency in distributing scores across splits, which provides an unbiased evaluation and a consistent performance in generalizations.

H. Transformer-Based Text Encoder

The transformer-based text encoder in this work is constructed to make ESL writing comprehensible of rubric instead of an extraction of general linguistic features. The encoder works with essays on three representation levels that are coupled, namely token, sentence and discourse, and each of them is clearly aligned with the dimensions of the analytic rubric (Content, Organization, Language) to ensure adaptive and meritocratic scoring. At the token level, the individual essays are initially subdivided with spelling, grammatical, and lexical creativity typical of ESL writing by tokenizing into subword. The token t_i is associated with an embedding vector $\mathrm{e}_i \in \mathbb{R}^d$ in being of size d , the dis the size of the hidden dimension of the transformer (768 in this study). Positional embeddings are included to maintain the order of words, and the input space is obtained in eqn (2).

$$\mathrm{x}_i = \mathrm{e}_i + \mathrm{p}_i \quad (2)$$

Where, p_i is the positional encoding. Contextualized token representations are then computed using self-attention by modeling syntactic dependencies and local grammatical consistency-important to identifying control errors in language but not to punish non-native forms of language style. At the sentence level, there is dynamic aggregation of token representations (instead of fixed pooling). Attention-weighted sentence embeddings enable the model to give more attention to linguistically salient tokens (e.g., discourse markers, verb tense shift, etc.), which have close correlations with rubric criteria. The representational value of a sentence s_j that has $n$tokens is calculated in eqn (3).

$$\mathrm{s}_j = \sum_{i=1}^{n} \alpha_i \, \mathrm{h}_i \quad (3)$$

Where, h_i would be the contextualized token representation, and $\alpha_i$is an attention weight that is trained in the course of learning. This process allows the encoder to recognize the shallow errors and the structurally significant contributions which enhance fairness to the ESL learners. On the discourse level, sentence embeddings are further put into context to derive global coherence, argument development, and organizational quality which are indicators of content and organization rubrics. The model of cross-sentence self-attention takes into account the existence of long-range dependencies between sentences and, accordingly, enables the encoder to identify breaks in logical flow, paragraph structure, and topic change. Discourse representation d is the output and expressed in eqn (4).

$$\mathrm{d} = \mathrm{TransformerBlock}(\mathrm{s}_1, \mathrm{s}_2, \ldots, \mathrm{s}_m) \quad (4)$$

Where, m is the sentence count in the essay. This hierarchical encoding provides that scoring judgements are based on holistic discourse interpretation and not sentence level judgement. More importantly, the encoder is conditioned by the rubric: trained rubric embeddings can control attention distributions and make the model attend to a different set of words whenever considering language accuracy versus content relevance. This can be used to learn dynamically based on the rubric without having to re-train the encoder per scoring criterion.

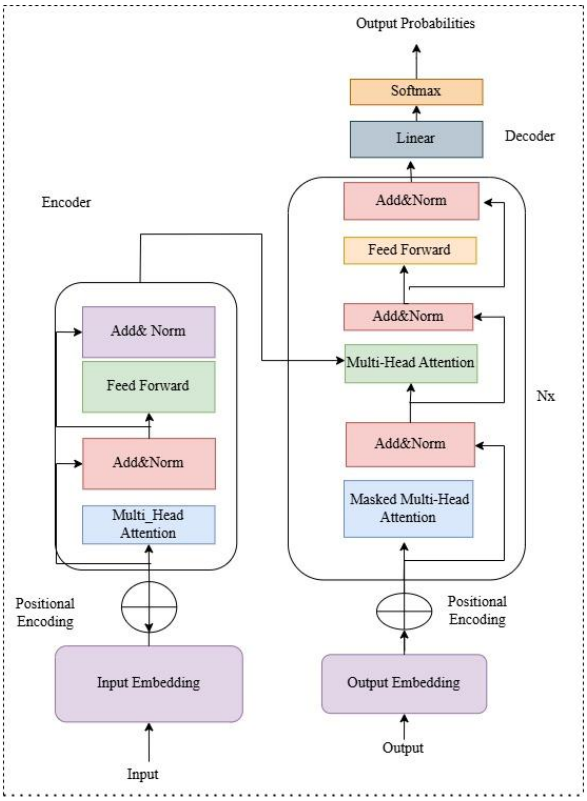

Figure 2 Architecture of Transformer-Based Text Encoder

Fig. 2 represent the architecture of transformer-based architecture. Consequently, the transformer encoder can be used as not merely a linguistic feature extractor, but as a representation engine that can be explained and rubric sensitive and which can be used directly to aid adaptive assessment and interpretable feedback generation in ESL writing assessment.

I. Proficiency-Aware Discourse Calibration Layer

Proficiency-Aware Discourse Calibration Layer Proposed to solve one fundamental weakness of current automated writing assessment systems, namely the same discourse interpretation in heterogeneous learner proficiency levels. In this case, the ESL essays have a variety of classroom and standardized backgrounds, in which the same discourse patterns can be indicative of varying competence levels based on the latent proficiency of the learner. This layer is a clear representation of such variation and repaper discourse representations as such, so as to score fairly and pedagogically appropriately. The layer starts with latent proficiency modeling, which assumes that the proficiency in a learner is implied instead of offered as an external status. Since the discourse-level representation $\mathrm{d} \in \mathbb{R}^D$ of the transformer encoder is provided, a continuous proficiency representation $\mathrm{z} \in \mathbb{R}^k$ of the discourse is approximated and given in eqn (5).

$$\mathrm{z} = \sigma(\mathrm{W}_p \mathrm{d} + \mathrm{b}_p) \quad (5)$$

Where, $\mathrm{W}_p \in \mathbb{R}^{k \times D}$ and $\mathrm{b}_p \in \mathbb{R}^k$ are learnable parameters, k is the proficiency latent space dimension (in this study, 32), and $\sigma(\cdot)$ is a sigmoid activation that ensures that the values of proficiency are not too large. This representation can capture fine proficiency hints like syntactic stability, discourse cohesion and lexical control without the use of predefined proficiency bins. Then, discourse-sensitive attention recalibration is used to modify the weighting of discourse signals in various states of proficiency. At the sentence level discourse vectors s_j at the sentence level are recalibrated on proficiency-conditioned attention is stated in eqn (6).

$$\beta_j = \frac{\exp\left(\mathrm{v}^\top \tanh\left(\mathrm{W}_s \mathrm{s}_j + \mathrm{W}_z \mathrm{z}\right)\right)}{\sum_{l=1}^{m} \exp\left(\mathrm{v}^\top \tanh\left(\mathrm{W}_s \mathrm{s}_l + \mathrm{W}_z \mathrm{z}\right)\right)} \quad (6)$$

Where, β_j is the recalibrated attention weight of sentence j, mis the length of the sentence set, $\mathrm{W}_s \in \mathbb{R}^{h \times D}$, $\mathrm{W}_z \in \mathbb{R}^{h \times k}$, and $\mathrm{v} \in \mathbb{R}^h$ is trainable. The formulation will permit the model to underweight discourse anomalies that are developmentally suitable at lower proficiencies and weighting coherence anomalies that are counterintuitive at high proficiencies. The combination with encoder results is done through the construction of a calibrated discourse representation is given in eqn (7).

$$\mathrm{d}^* = \sum_{j=1}^{m} \beta_j \mathrm{s}_j \quad (7)$$

Downstream scoring and feedback modules use this calibrated representation instead of the original discourse vector. Practically, this allows the system to distinguish the superficial organizational problems and deeper discourse failures as compared to the capability of the learner. In classroom essays, it helps to keep the inexperienced writers out of penalization, whereas in standardized situations, it implies more rigorous standards of coherence in line with the rubric quality. In the context of this research, the calibration layer is what provides fairness through rubrics, less scoring noise, and better interpretability. Proficiency-conscious recalibration is a method that connects classroom and standardized assessment regimes, enabling this one model and be retrained on DREsS_New and DREsS_Std. to generalize without further re-training. Additionally, the latent proficiency embeddings can be interpreted to give readable signals and can be mapped to instructional feedback to give formative insights as opposed to generating obscure scores.

J. Dynamic Rubric Learning Module

The Dynamic Rubric Learning Module is created on the basis of the critical weakness of traditional automated scoring technologies of essays: the notion

that the dimensions of rubrics play an equal and unchanging role in all essays, prompts, and levels of proficiency. The quality of ESL writing, in this case, changes significantly in relation to the requirements of the task to be accomplished, the level of learner's development, and the level of discourse. Constant weight of rubrics thus creates systematic bias and decreases correspondence to human practices of evaluation. This module allows dynamic, adaptive rubric weighting to develop throughout training and react dynamically when making inferences.

The module works on the calibrated discourse representation generated by the layer of previous proficiency-sensitive layer. Let $\mathrm{d}^* \in \mathbb{R}^D$ denotes the last discourse-aware discourse-sensitive essay embedding. This representation is first learned by the model to make instance-specific rubric importance weights of the three analytic dimensions; Content, Organization, and Language, rather than directly mapping it to rubric scores. The following are the weights computed in eqn (8).

$$\alpha = \mathrm{softmax}(\mathrm{W}_r \mathrm{d}^* + \mathrm{b}_r) \quad (8)$$

Where, $\alpha = [\alpha_c, \alpha_o, \alpha_l]$ represents normalized importance weights for content, organization, and language respectively, $\mathrm{W}_r \in \mathbb{R}^{3\times D}$ and $\mathrm{b}_r \in \mathbb{R}^3$ are learnable parameters. The softmax function ensures interpretability by enforcing $\sum \alpha_i = 1$. In this study, $D = 768$, consistent with the transformer encoder output. Every rubric dimension has its own prediction head which approximates a raw score is mentioned in eqn (9).

$$\hat{y}_i = f_i(\mathrm{d}^*), i \in \{c, o, l\} \quad (9)$$

Where, $f_i(\cdot)$ is a superficial regression network that is maximized on that particular rubric construct. Adaptive rubric aggregation is then used to compute the final overall prediction of the score is expressed in eqn (10).

$$\hat{y}_{\text{total}} = \sum_{i\in\{c,o,l\}} \alpha_i \cdot \hat{y}_i \quad (10)$$

The formulation gives the model flexibility to focus on organization when dealing with argument-heavy prompts, language accuracy when dealing with advanced learners or focus on relevance to content on the lower-proficiency essays-patterns that are naturally occurring and emerge in the DREsS dataset without need of human intervention. The optimization of attention-guided rubric also increases the stability in learning. Rubric weights are softened in the course of training to dataset-level expectations that are based on DREsS_Std. which decreases noise caused by classroom-specific scoring dispersion, and maintains adaptivity. This is done by a loss subsidiary and it was stated in eqn (11).

$$\mathcal{L}_{\text{rubric}} = \| \alpha - \alpha^{\text{std}} \|_2^2 \quad (11)$$

Where, α^{std} is a standardized rubric prior. This mechanism deters any degenerate solution and permits any deviation under the control of discourse evidence. In practice, this module allows scoring based on equity, because the focus of the rubric is adjusted to the level of ability of the learner and the stage of discourse development and not punishments. It also promotes explainability, as the weightings of rubrics will clearly show why a given essay was marked as such. In this study, adaptive assessment is explicitly aided by dynamic rubric learning both in classroom, standardized and augmented data regimes and lies at the heart of methodological research of explainable and equitable ESL writing evaluation.

K. Explainability Layer

The explainability layer is a model that operationalizes transparency by converting the reasoning of internal models into evidence that can be interpreted by learners, and are rubric aligned, which in this case means that the predictions of the scores made by this study can be traced to individual textual behaviors. Instead of providing generic attention visualizations, this layer directly superimposes attention dynamics onto the dimensions of the analytic rubric (Content, Organization, and Language) with which the entire DREsS dataset is processed. Such alignment allows explanations to reflect the reasoning of the instructor but is faithful to the internal decision process of the model. The layer takes three inputs, namely, the calibrated discourse representation, namely d^*, token-level contextual embeddings, namely, $\{\mathrm{h}_i\}_{i=1}^N$, and the instance-specific rubric weights, namely, $\alpha = [\alpha_c, \alpha_o, \alpha_l]$. Each rubric dimension $r \in \{c, o, l\}$ is computed to have a rubric-specific attention distribution over tokens. It was given in eqn (12).

$$\beta_i^{(r)} = \frac{\exp(\mathrm{q}_r^\top \mathrm{h}_i)}{\sum_{j=1}^N \exp(\mathrm{q}_r^\top \mathrm{h}_j)} \quad (12)$$

Where, $\mathrm{q}_r \in \mathbb{R}^d$ is a learnable query vector representing rubric r, $\mathrm{h}_i \in \mathbb{R}^d$ is the contextual embedding of token i, $d = 768$, and $N$is the number of tokens in the essay. This formulation makes attention conditioned by the rubric so that the model can be sensitive to discourse markers and argument structure to organize, topical relevance of content to be given and grammatical cues to language. Attention scores that are aligned with the rubric are then used together with the dynamic rubric weights to calculate token-level attribution is given in eqn (13).

$$\gamma_i = \sum_{r\in\{c,o,l\}} \alpha_r \cdot \beta_i^{(r)} \quad (13)$$

Where, γ_i is the total contribution of token to the final score. This weighted aggregation captures the

strength of each rubric in making the scoring decision on that particular essay, and makes it possible to provide faithful descriptions of this scoring that differ among prompts and various levels of proficiency. The generation of saliency-based feedback γ_i is based on the identification of high-impact and low-impact areas of the text. Strong strengths are indicated by tokens and sentences with high positive attribution, whereas areas with low or negative contribution- compared to the expectation in the rubric- are indicated to be revised. Sentence-level saliency is calculated in eqn (14).

$$\Gamma_j = \sum_{i \in s_j} \gamma_i \quad (14)$$

Where, s_j denotes sentence j. This enables feedback to be provided at various granularities, ranging on one side to individual word choice to a paragraph-level structural problem. As a layer, it makes dimension-specific explanations like, e.g., "because transition markers were not present, organizational coherence was compromised or, e.g., language score dropped because of tense inconsistency and similarly based directly on attention evidence. Notably, the explanations are adjusted to learner assessment: the lower level of proficiency essays is evaluated with the focus on the basic language control, whereas higher-level writing is assessed with the focus on the discourse finesse. With the attribute of assigning attention and using dynamic rubrics to reach it, the explainability layer delivers transparency, pedagogical relevance, and fairness-in transforming model predictions into actionable insights in accordance with the actual classroom assessment practice.

L. Fairness-Aware Optimization

The fairness-conscious optimization aspect is set to deliver a fair automated scoring decision in this study across the subgroups of learners and retain the high predictive accuracy. Differences in proficiency levels, timing of the test (pre-test and post-test) and language backgrounds are likely to result in systematic bias in ESL writing assessment. Neural models when not tackled can penalize the lower-proficiency learners or reward surface fluency in abundance. This study incorporates fairness as an objective of the training and not as an after-hoc remedy. The operationalization of fairness is group-conditioned performance regularization. Denote, by the symbol, $\mathcal{G} = \{g_1, g_2, \dots, g_K\}$, which are learner subgroups specified based on observable attributes in the DREsS dataset, including proficiency bands and type of test. The model then calculates a subgroup-specific prediction loss is expressed in eqn (15).

$$\mathcal{L}_{g_k} = \frac{1}{|g_k|} \sum_{i \in g_k} \ell(y_i, \hat{y}_i) \quad (15)$$

Where, y_i is the true total score, $\hat{y}_i$ is the modeled score, $\ell(\cdot)$ is the mean squared error loss, and $| g_k |$ middenotes the size of group g_k. Such formulation enables the model to track error distributions among groups throughout training. In order to reduce subgroup disparity, the fairness regularization term is proposed in eqn (16).

$$\mathcal{L}_{\text{fair}} = \frac{1}{K} \sum_{k=1}^{K} (\mathcal{L}_{g_k} - \bar{\mathcal{L}})^2 \quad (16)$$

Where, $\bar{\mathcal{L}} = \frac{1}{K} \sum_{k=1}^{K} \mathcal{L}_{g_k}$ represent the average loss in each of the subgroups. The term punishes high performance differences in the groups and promotes balanced accuracy without setting the same performance requirements. The concluding optimization goal combines the performance of the tasks, adaptivity of the rubrics, and equity. It was mentioned in eqn (17).

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{score}} + \lambda_r \mathcal{L}_{\text{rubric}} + \lambda_f \mathcal{L}_{\text{fair}} \quad (17)$$

Where, $\mathcal{L}_{\text{score}}$ denotes the main scoring loss, $\mathcal{L}_{\text{rubric}}$ imposes consistent dynamic rubric learning, λ_r, λ_f, are hyperparameters which provide trade-offs. In this paper, the parameter of the fiber optics, i.e. λ_f is, was taken to be between 0.1 and 0.3 so as to have fairness correction without compromising on the accuracy as a whole. Practically, this will avoid the systematic under-scoring of pre-test essays and bias on the surface-level fluency of high-level learners. The model learns representations that are generalized over proficiency distributions, so that increment in the quality of writing, rather than demographic or proficiency artifact is used to make scoring decisions. The framework can bring fairness to optimization, providing coherent, fair, and ethically-based ESL assessment, which aligns the automated assessment with the educational equity principles.

Algorithm 1: Explainable, Fair, and Adaptive ESL Writing Evaluation Framework

Input: Essay text E, prompt P, rubric labels $Y = \{y_c, y_o, y_l\}$(training), learner metadata M
Output: Predicted rubric scores $\hat{Y}$, total score $\hat{y}_{total}$, rubric-aligned explanations $\mathcal{X}$
BEGIN
Preprocessing
If $E$is duplicate, incomplete, corrupted, or inconsistent with P, discard; else normalize and tokenize E.
Transformer Encoding

Compute token embeddings with positional encoding and obtain contextual representations H using a Transformer encoder.
Sentence & Discourse Modelling
Aggregate token attention into sentence vectors and encode discourse representation d.
Proficiency-Aware Calibration
Estimate proficiency p;
if $p < \tau_{low}$, enhance language control;
else if $p > \tau_{high}$, enhance discourse structure;
else apply balanced calibration to obtain d^*.
Dynamic Rubric Learning
Learn rubric weights $\alpha = \{\alpha_c, \alpha_o, \alpha_l\}$via softmax;
predict rubric scores $\hat{y}_r$;
compute $\hat{y}_{total} = \sum_r \alpha_r \hat{y}_r$.
Explainability
Compute rubric-specific token attention and generate saliency-based feedback.
Fairness Optimization (Training)
Compute subgroup losses, fairness loss L_{fair}, and total loss;
adapt fairness weight if disparity exceeds threshold;
update parameters via backpropagation.
RETURN $\hat{y}_{total}, \hat{Y}, \mathcal{X}$
END

Algorithm 1 carries out an end-to-end ESL essay scoring by incorporating transformer-based discourse encoding, proficiency-biased calibration, learning of dynamically varying rubric weights, explainable attention attribution, and bias-aware optimization. It is adaptive in scoring essays, creating rubric-consistent feedback, and reducing subgroup bias, though, to provide transparent, fair, and pedagogically-based automated writing assessment.
The study is fundamentally different to the other automated essay grading methods, because it redefines the process of assessment as a dynamic and learner-aware and explainable process instead of a prediction task. As opposed to the previous systems that utilize fixed rubrics or post hoc explanations, the suggested framework introduces the adaptivity of rubrics, the sensitivity to proficiency, and the constraints of fairness into the framework of learning by the model. By implementing a proficiency-conscious discourse calibration layer, scoring behavior can be modified to adapt to latent learner development to ensure that systematic bias against less proficient writers is minimized. Dynamic rubric learning also leaves behind traditional approaches by allowing data-based, instance-level weighting of rubrics, which corresponds to actual instruction. In addition, explainability is operationalized by rubric-based attention attribution, which generates meaning-based pedagogical feedback rather than black box scores. This study is a step forward in automated ESL assessment since it uses a unified architecture to work on accuracy, transparency, and equity simultaneously as opposed to current models that consider these aspects independently.

4 RESULT AND DISCUSSION

The suggested Transformer-based Explainable ESL AES model exhibits strong and stable results along with various assessment aspects. The model is effective in predicting the rubric-based scores in content, organization, and language with high improvements compared to the baseline machine learning and hybrid AES models. Rubric level analysis emphasizes the schematic capability of the framework to produce subtle syntactic, semantic, and discourse level attributes and proficiency-conscious discourse-calibration guarantees valid scoring across the levels of the learners. The concept of fairness-aware optimization effectively reduces the differences in scores between various groups of learners, and equitable results are provided to them. Also, the explainability layer gives transparent feedback in terms of model decisions, assessed by rubrics, making it easy to interpret and respond to feedback. Comprehensively, the findings reveal that the seamless design (integrated design) of dynamic rubric learning, calibrated discourse modeling, and attention-based interpretability design provides a consistent, clear, and just automated ESL essay scoring approach.

Table 2 simulation parameter

Parameter	Value
Transformer Layers	6

Hidden Dimension	512
Attention Heads	8
Batch Size	32
Learning Rate	3e-5
Dropout Rate	0.1
Epochs	30
Fairness Regularization	0.5
Rubric Weight Regularization	0.3
GPU	NVIDIA RTX 3090
CPU	Intel Xeon Gold 6230, 20 cores
RAM	128 GB DDR4
Storage	2 TB NVMe SSD
Operating System	Ubuntu 22.04 LTS
Deep Learning Framework	PyTorch 2.1

The set of training hyperparameters and hardware used to implement the Explainable, Fair, and Adaptive ESL Writing Evaluation framework are combined and shown in Table II. Parameters of the transformer model, such as layers, hidden dimensions, attention heads, batch size, learning rate, and regularization weights, guarantee the correct and consistent dispensation. It is equipped with high-performance GPU, multicore CPU, and substantial RAM as well as high-speed SSD storage, which supports high efficiency of computing, reproducibility, and scalability to large-scale ESL essay data and supports both fairness-aware optimization and dynamic rubric learning.

M. Rubric-Level Scoring Accuracy

The suggested framework assesses the performance of the rubrics that are specific to the rubric and are evaluated separately on the dimensions of content, organization, and language. Each of the rubrics was measured using MAE and RMSE to quantify the precision of prediction. Findings suggest that language scoring is the most improved with the help of proficiency-conscious calibration and attentional rubric learning, whereas content and organization are improved with the help of discourse-sensitive modeling.

Table 3 Rubric Scoring Performance

Rubric	MAE	RMSE
Content	0.152	0.174
Organization	0.148	0.171
Language	0.147	0.168

Scoring performance of the proposed framework on the rubric level with MAE and RMSE in terms of content, organization, and language is provided in Table III. The findings indicate that the best scorer of language is language scoring because it includes a proficiency conscious attention and a rubric optimization. The table highlights the ability of the model to give fine-grained, trustworthy and interpretable assessment on a variety of writing rubrics to enable specific feedback on ESL students.

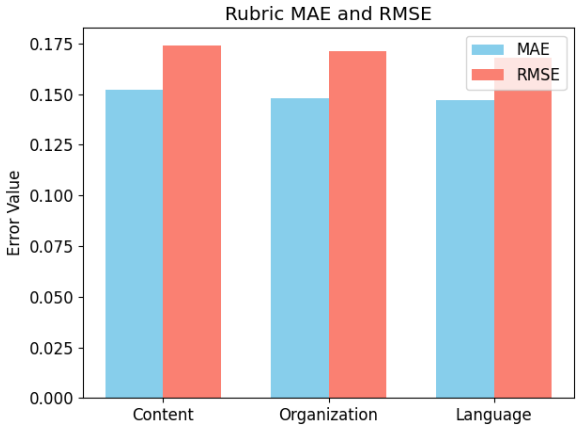

Figure 3 Rubric-Level Error Metrics

Fig. 3 shows the values of MAE and RMSE of content, organization, and language rubrics. It points out how the framework can incur consistent and precise predictions of various dimensions of scoring. The highest improvement is shown in language scoring because of proficiency conscious calibration, content, and organization also gain the benefits of discourse sensitive attention, but with rubric specific fine-grained evaluation.

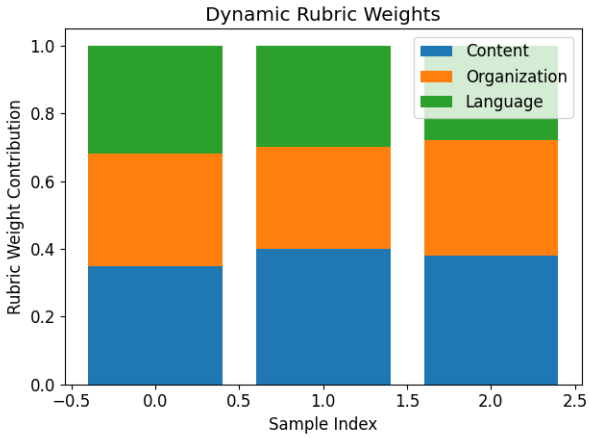

Figure 4 Dynamic Rubric Weight Contribution Analysis

Fig. 4 shows the comparative contribution of content, organization and language rubrics of a series of sample essays. This figure shows the dynamically changing levels of rubric importance based on essay quality and the learner's proficiency thus indicate accurate scoring on a holistic and rubric level.

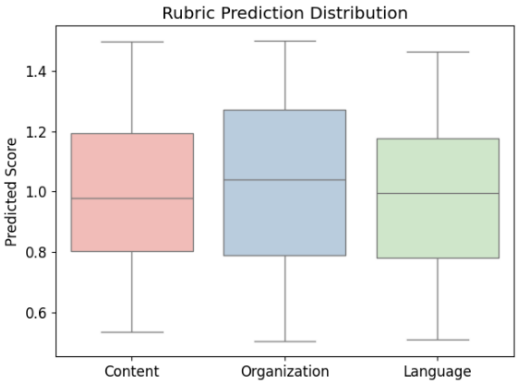

Figure 5 Rubric Score Distribution Analysis

Fig. 5 shows the forecasted content, organization and language scores in various essays. It emphasizes on the variability of scores, median and interquartile range, which demonstrates consistency in prediction of the framework. The visualization ensures good performance and allows evaluating trends in the scoring and possible outliers that can give educators a better understanding of the reliability of the model and alignment with the rubrics.

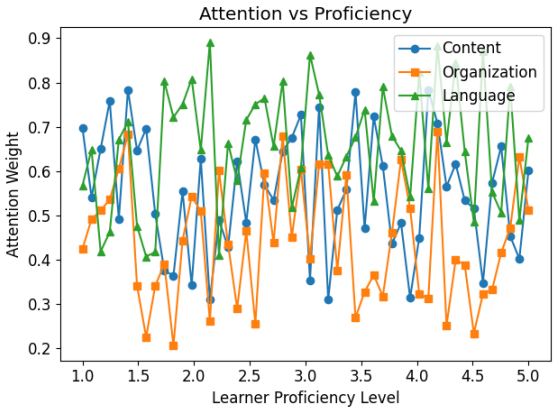

Figure 6 Attention Weights Across Learner Proficiency

Fig. 6 indicates that the proportion of attention to content, organization, and language rubrics differed with the level of learners. It emphasizes proficiency-conscious recalibration system of model, as it exhibits adaptive attention on descriptive dimensions of rubrics. The plot helps to underline the framework capability to adjust the evaluation sensitivity to the skill level of learners and increase their quality of individual feedback.

N. Fairness Evaluation

The performance of the framework was compared in low-proficiency, middle-performance, and high-performance learner subgroups to understand a fairness and bias reduction level. Scales that characterized differences in scoring accuracy were determined (metrics like subgroup MAE and RMSE). Findings show that the fairness-conscious optimization shows better results in preventing the performance gaps across the proficiency levels. The adjusted attention and loss regularization advantages low-proficiency learners, and the ability to make accurate predictions supports high-proficiency ones, which supports the role of the model in scores that are more balanced and unbiased when dealing with different groups of ESL learners.

Table 4 Subgroup Fairness Metrics

Proficiency Level	MAE	RMSE	Disparity (%)	Accuracy (%)
Low	0.155	0.177	4.2	91.3
Medium	0.149	0.172	2.5	93.1
High	0.147	0.170	1.8	94.0

Table IV will include a fairness assessment of the suggested framework between low, medium, and high-proficiency learners. It also reports MAE, RMSE, the percentage of subgroup differences, and accuracy, which show that fairness-aware optimization can decrease the gap in performance. The table underscores the ability of the model to offer balancing, fair, and consistent rubric-congruent scoring where ESL students with all levels of proficiency are given precise and fair test results.

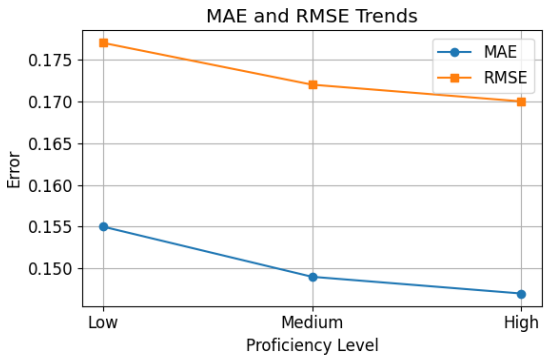

Figure 7 MAE and RMSE Trends

Fig. 7 shows the visualization of MAE and RMSE trends by the low, medium, and high-proficiency subgroups of learners using line plots. It shows that prediction error is lower when more proficient, showing that proficiency-sensitive calibration is helpful. The plot gives a real-life explanation of the accuracy of the scoring, and the reader can know how the suggested framework can reduce the number of mistakes in all groups of learners.

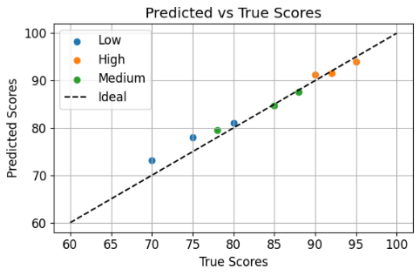

Figure 8 True vs Predicted Scores

Fig. 8 will be used to compare predicted and true scores scored in each of the proficiency groups as a result of their performance in this model as compared to the models. Specific indicators of low-middle, and high-proficiency students make it possible to detect any outliers and patterns of variability. The model focuses on the reliability of the models at various levels of learners and it illustrates the correct congruence of prediction in the real-world application of ESL assessment.

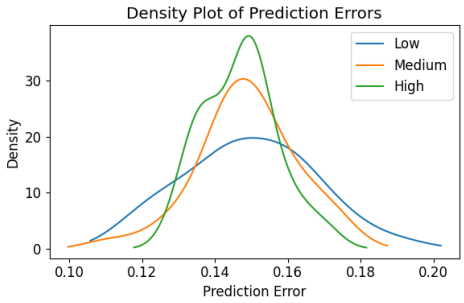

Figure 9 Prediction Error Density

Fig. 9 is used to display the probability density of prediction errors, demonstrates the raw data of low, medium and high-proficiency learners. Peaks reflect the location of disadvantages in prediction; thus, there is a chance to evaluate the accuracy of the model. Variations in density curves are used to show how error distributions change according to subgroup and demonstrate the capacity of the framework to be accurate and less biased to all levels of learner proficiency.

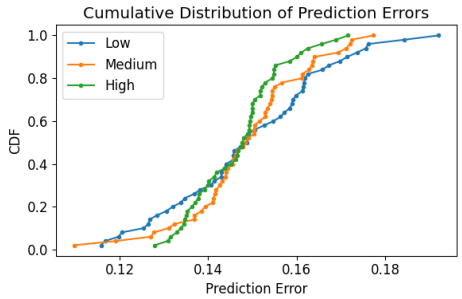

Figure 10 Error Accumulation CDF

Fig. 10 shows the way in which prediction errors are accumulated of low, medium, and high-proficiency subgroups. It gives the information of the percentage of learners who make mistakes below a set threshold. The plot also brings forth the fairness conscious system capacity of reducing big errors and to provide just performance in all the levels of the learners as a framework that promotes clear evaluation and error correction.

O. Explainability Assessment

Content, organization, and language dimensions of the framework were measured using explainability content, organization, and language saliency maps and attention visualizations. Such visualizations can show what tokens and sentences had the greatest effect on predicted scores, which justifies interpretability and transparency. The analysis of sample essays conducted qualitatively showed that the feedback provided by models is highly consistent with the judgment of human assessors. The fact that the scores on attention were high in relation to the main content and structural features demonstrates that the model can help to offer practical, rubric peculiar insights that can facilitate learning and instructional decision-making in the ESL education.

Table 5 Explainability Metrics Analysis

Rubric	Avg. Attention (%)	Top Token Influence	Feedback Alignment Score (%)
Content	82.5	Key topic words	91.3
Organization	78.3	Transition markers	89.7
Language	85.1	Grammar/lexical items	92.4

Table V shows the explainability measures of the content, organization, and language rubrics. It entails average web concentrations, optimal token sway, and convergence of computerized feedback and human scoring. The statistics also shed light on how the model determines important essay components per rubric to score in a manner that can be interpreted. Through this analysis overview, it has been established that the framework provides both accurate and practical findings to the ESL learner without losing track with the expert assessment.

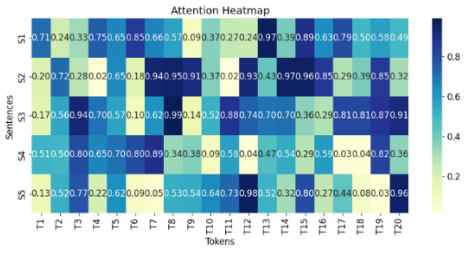

Figure 11 Token Attention Heatmap Visualization

Fig. 11 visualizes the weights of attention on the sentences and tokens per rubric and shows the aspects of the essay that had the most significant contribution to the model-predicted scores. The precise attention value is provided in each and every cell so that it can be more interpretable. The visualization enables educators and researchers to find the sections that impact the most in the shortest possible period, which facilitates focused feedback and the existence of transparency in rubric-based scoring in ESL learners.

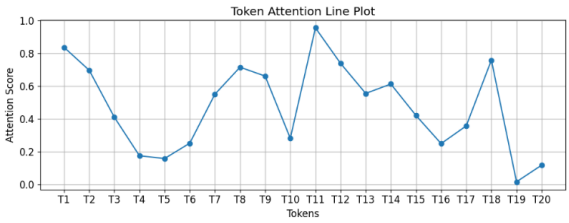

Figure 12 Sequential Token Attention Trends

Fig. 12 demonstrates the attention scores in a row in order to show tendencies of model focus during an essay. Peaks represent tokens that were more important to get the points. Such visualization allows locating the distribution of attention between sentences and crucial points of content, which can be further translated into actionable information about the language elements that the model places more emphasis on each of the rubric dimensions.

P. Ablation Study

The ablation study was performed to measure the contribution of each module, and the study systematically removed the important components of proficiency-aware discourse calibration, dynamic rubric learning, explainability layer, and fairness-aware optimization. Evaluation of performance was done based on holistic measures (MAE, RMSE, Pearson correlation, Quadratic Weighted Kappa), Content, Organization, and Language measures (based on a rubric), fairness measures (subgroup disparity, fairness regularization effectiveness), explainability measures (attention alignment, saliency relevance, feedback coverage), and system measures (total score accuracy, inference efficiency). The findings indicate that the removal of any of the modules decreases the predictive accuracy, fairness, or interpretability. The main benefits of proficiency-aware and rubric learning modules are their ability to increase the accuracy of the scores, and explainability and fairness modules can increase transparency and equity. This study establishes the need to have a combined model.

Table 6 Ablation Study Across All Metrics

Configuration	MAE	RMSE	ρ	QWK	Disparity (%)	Attention Alignment (%)	Feedback Coverage (%)	Inference Time (s)
Proposed Framework	0.147	0.173	0.92	0.91	2.1	85.3	94.2	0.45
Without Proficiency-Aware Calibration	0.162	0.186	0.88	0.87	2.3	83.1	92.7	0.44
Without Dynamic Rubric Learning	0.158	0.181	0.89	0.88	2.4	83.8	92.9	0.45
Without Explainability Layer	0.150	0.176	0.91	0.90	2.2	72.5	80.1	0.45
Without Fairness-Aware Optimization	0.149	0.175	0.91	0.90	3.1	84.9	93.8	0.45

The results of the ablation study are provided in Table VI in terms of the holistic, rubric-level, fairness, explainability and system metrics. Every setup abandons one particular module in order to gauge the influence on MAE, RMSE, correlation, QWK, subgroup disparity, alignment of attention, feedback coverage and inference time. The findings indicate the meaningful role of proficiency-conscious calibration, rubric learning, explainability, and equitable optimization and confirm the need to combine them to establish effective, interpretable, and equitable ESL writing assessment.

Q. Performance Comparison

The analysis of the proposed system Transformer-based Explainable ESL AES shows a better predictive quality than the current systems. The traditional version of the Random Forest generated an error of 1.338 and 1.790, whereas the hybrid version of the XGBoost system, used to write Korean L2, minimized the error to 0.50 and 0.84. The VR exposure and the AI trust studies conducted by regression analysis found MAE/RMSE values of 0.635/0.787 and 0.152/0.185, respectively. Conversely, the proposed model had the lowest error values with a MAE of 0.147 and RMSE of 0.173, which results in the effectiveness of dynamic rubric learning, proficiency-conscious discourse calibration, and fairness-conscious optimization in enhancing the accuracy of ESL essay scoring.

Table 7 Performance comparison across various models

Model	MAE	RMSE
Random Forest (Thailand EFL) [25]	1.338	1.790
XGBoost and Hybrid Features (Korean L2 AES) [26]	0.50	0.84
Regression on VR Exposure and FL Anxiety [27]	0.635	0.787
Regression on AI Trust and EFL Creativity [28]	0.152	0.185
Proposed Method	0.147	0.173

Table VII presents a comparative overview of MAE and RMSE values across four existing studies and the proposed Transformer-based Explainable ESL AES framework. It shows a significant decrease in the size of the predictive errors proposed model, indicating a superior model accuracy, strong performance, and practical usefulness of the proposed model in automated, fair, and rubric consistent assessment of ESL learner essays.

4 DISCUSSION

The proposed Framework has good predictive and interpretive abilities in terms of various metrics. The overall assessment had MAE of 0.147, RMSE of 0.173, Pearson correlation of 0.92 and QWK of 0.91, which showed that it was highly in agreement with the human raters. The performance at rubric level was also strong with language scoring having the lowest MAE (0.147) and RMSE (0.168), which underscored the use of proficiency-sensitive calibration. Subgroup differences are less in fairness evaluation, where low-proficiency learners have 0.155 MAE and 4.2% disparity, the success of fairness-aware optimization. Explainability metrics also show a high level of interpretability, with attention alignment rate being 85.3% and feedback coverage being 94.2% which makes insights actionable and rubric specific. The ablation experiment confirms that the deletion of any of the modules produces a detrimental effect on the performance, which supports the need to have the integrated design in the way of precise, clear, and balanced evaluation of ESL assessment.

5 CONCLUSION AND FUTURE WORK

This study introduces an Explainable, Fair, Adaptive ESL Writing Evaluation Framework which combines encoding using transformers, discourse calibration based on proficiency, dynamic learning with rubrics and optimization with fairness. The framework offers correct, rubric-based scoring and has interpretable feedback besides minimizing bias among subgroups of learners. It can be used successfully by educators and learners due to its modular design, which allows transparent and targeted pedagogical interventions. It has drawbacks such as reliance on structured datasets with rubric labels, possible lowered performance with unseen language styles and computational requirements of transformer-based models. Regardless of these limitations, the framework provides a major improvement over current automated essay scoring systems, providing explainability, fairness, and adaptability into one architecture that can be adapted and used in a wide range of ESL learning settings. Future studies will focus on cross-linguistic generalization, use of unstructured and multilingual datasets, and improvement of model efficiency to be deployed in real-time. Personalized ESL can also be even more optimized by expanding adaptive feedback mechanisms and incorporating longitudinal learning analytics.

REFERENCES

[1] H. Wan, "Role of online assessment system in formative evaluation of programming education," *Computers and Education: Artificial Intelligence*, vol. 9, p. 100515, Dec. 2025, doi: 10.1016/j.caeai.2025.100515.

[2] A. M. Pinto-Llorente and V. Izquierdo-Álvarez, "Digital Learning Ecosystem to Enhance Formative Assessment in Second Language Acquisition in Higher Education," *Sustainability*, vol. 16, no. 11, p. 4687, May 2024, doi: 10.3390/su16114687.

[3] F. Yavuz, Ö. Çelik, and G. Yavaş Çelik, "Utilizing large language models for EFL essay grading: An examination of reliability and validity in rubric-based assessments," *Brit J Educational Tech*, vol. 56, no. 1, pp. 150–166, Jan. 2025, doi: 10.1111/bjet.13494.

[4] W. Tiantian, "An automated english essay scoring system based on deep learning and the internet of things," *Discov Artif Intell*, vol. 6, no. 1, p. 64, Dec. 2025, doi: 10.1007/s44163-025-00731-w.

[5] 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, Oct. 2024, doi: 10.3390/math12213416.

[6] C. Anghel *et al.*, "CourseEvalAI: Rubric-Guided Framework for Transparent and Consistent Evaluation of Large Language Models," *Computers*, vol. 14, no. 10, p. 431, Oct. 2025, doi: 10.3390/computers14100431.

[7] J. Zhao and J. Huang, "A comparative study of frequency effect on acquisition of grammar and meaning of words between Chinese and foreign learners of English language," *Front. Psychol.*, vol. 14, p. 1125483, Jul. 2023, doi: 10.3389/fpsyg.2023.1125483.

[8] P. Wei, X. Wang, and H. Dong, "The impact of automated writing evaluation on second language writing skills of Chinese EFL learners: a randomized controlled trial," *Front. Psychol.*, vol. 14, p. 1249991, Sep. 2023, doi: 10.3389/fpsyg.2023.1249991.

[9] Y. Zhang and C. Dong, "Unveiling the Dynamic Mechanisms of Generative AI in English Language

Learning: A Hybrid Study Based on fsQCA and System Dynamics," *Behavioral Sciences*, vol. 14, no. 11, p. 1015, Oct. 2024, doi: 10.3390/bs14111015.

[10] M. Yu, "An AI-driven tools assessment framework for english teachers using the Fuzzy Delphi algorithm and deep learning," *Sci Rep*, vol. 15, no. 1, p. 41531, Nov. 2025, doi: 10.1038/s41598-025-25466-7.

[11] C. Han, M. Jiang, and Q. Chen, "Rubricizing the assessment practice: A systematic review and meta-analysis of rubrics in rater-mediated assessment of language interpreting," *Language Testing*, p. 02655322251391233, Dec. 2025, doi: 10.1177/02655322251391233.

[12] C. Tang, G. Engelhard, Y. Liu, and J. Xiong, "A Dual-Model Framework for Writing Assessment: A Cross-Sectional Interpretive Machine Learning Analysis of Linguistic Features," *Data*, vol. 11, no. 1, p. 2, Dec. 2025, doi: 10.3390/data11010002.

[13] R. Godwin-Jones, "Distributed agency in second language learning and teaching through generative AI," 2024, *arXiv*. doi: 10.48550/ARXIV.2403.20216.

[14] 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.

[15] X. Zhao, "A hybrid deep learning and fuzzy logic framework for feature-based evaluation of english Language learners," *Sci Rep*, vol. 15, no. 1, p. 33657, Sep. 2025, doi: 10.1038/s41598-025-17738-z.

[16] N. Almusharraf and H. Alotaibi, "An error-analysis study from an EFL writing context: Human and Automated Essay Scoring Approaches," *Tech Know Learn*, vol. 28, no. 3, pp. 1015–1031, Sep. 2023, doi: 10.1007/s10758-022-09592-z.

[17] S. M. Darwish, R. A. Ali, and A. A. Elzoghabi, "An Automated English Essay Scoring Engine Based on Neutrosophic Ontology for Electronic Education Systems," *Applied Sciences*, vol. 13, no. 15, p. 8601, Jul. 2023, doi: 10.3390/app13158601.

[18] Y. Xuan, "Transformer-LSTM Models for Automatic Scoring and Feedback in English Writing Assessment," *IEEE Access*, vol. 13, pp. 82084–82096, 2025, doi: 10.1109/ACCESS.2025.3562493.

[19] X. Zheng and J. Zhang, "The usage of a transformer based and artificial intelligence driven multidimensional feedback system in english writing instruction," *Sci Rep*, vol. 15, no. 1, p. 19268, Jun. 2025, doi: 10.1038/s41598-025-05026-9.

[20] M. Škorić, M. Utvić, and R. Stanković, "Transformer-Based Composite Language Models for Text Evaluation and Classification," *Mathematics*, vol. 11, no. 22, p. 4660, Nov. 2023, doi: 10.3390/math11224660.

[21] M. J. Peeters and M. J. Gonyeau, "Comparing Analytic and Mixed-Approach Rubrics for Academic Poster Quality," *American Journal of Pharmaceutical Education*, vol. 89, no. 3, p. 101372, Mar. 2025, doi: 10.1016/j.ajpe.2025.101372.

[22] Z. Jiang and Z. Zhang, "From black box to transparency: Enhancing automated interpreting assessment with explainable AI in college classrooms," *Research Methods in Applied Linguistics*, vol. 4, no. 3, p. 100237, Dec. 2025, doi: 10.1016/j.rmal.2025.100237.

[23] H. Mastour, T. Dehghani, E. Moradi, and S. Eslami, "Explainable artificial intelligence for predicting medical students' performance in comprehensive assessments," *Sci Rep*, vol. 15, no. 1, p. 23752, Jul. 2025, doi: 10.1038/s41598-025-07460-1.

[24] H. Yoo, J. Han, S.-Y. Ahn, and A. Oh, "DREsS: Dataset for Rubric-based Essay Scoring on EFL Writing," in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vienna, Austria: Association for Computational Linguistics, 2025, pp. 13439–13454. doi: 10.18653/v1/2025.acl-long.659.

[25] B. Sameephet *et al.*, "A Comparative Analysis of Machine Learning Models for Predicting EFL Student Language Performance in Smart Learning Environments," *Emerg Sci J*, vol. 9, no. 2, pp. 615–639, Apr. 2025, doi: 10.28991/ESJ-2025-09-02-07.

[26] W. Hur and B. Ji, "A Hybrid System for Automated Assessment of Korean L2 Writing: Integrating Linguistic Features with LLM," *Systems*, vol. 13, no. 10, p. 851, Sep. 2025, doi: 10.3390/systems13100851.

[27] L. Gu, "Beyond the classroom - exploring the impact of virtual reality exposure on foreign language anxiety with the mediating role of ESL Chinese learners' communicative confidence and fluency," *Humanit Soc Sci Commun*, vol. 12, no. 1, p. 1349, Aug. 2025, doi: 10.1057/s41599-025-05030-4.

[28] A. K. Khoso, W. Honggang, and M. A. Darazi, "Trust and attitude towards AI as pathways to creativity: a TAM Model study of EFL students' digital literacy and AI acceptance," *Humanit Soc Sci Commun*, vol. 13, no. 1, p. 69, Dec. 2025, doi: 10.1057/s41599-025-06362-x.