Natural Language Processing (2026), **32**, pp. 638–661
doi:10.1017/nlp.2026.10037

CAMBRIDGE UNIVERSITY PRESS

ARTICLE

Unleashing the potential of pre-trained language model in multi-dimensional writing assessment: towards more robust and explainable scoring

Xiaoyi Tang

School of Foreign Studies, University of Science and Technology Beijing, China
Email: tangxiaoyi@ustb.edu.cn

(Received 19 July 2025 UTC; revised 16 July 2026 UTC; accepted 16 August 2026 UTC;
first published online 22 September 2026 UTC)

Abstract

Rapid advancements in education assessment and natural language processing have fundamentally transformed the methodologies employed in multi-dimensional writing assessment. Neural network approaches have shown remarkable effectiveness, obviating the need for handcrafted feature engineering. However, the potential of pre-trained language models for constructing multi-dimensional automated essay scoring (AES) models still needs to be explored. The present study addresses critical challenges, including the resilience of AES models to subtle input perturbations leading to prediction inaccuracies and the elucidation of the intricate mechanisms within deep learning frameworks. We propose an integrated approach leveraging multi-task learning and the pre-trained model Bidirectional Encoder Representations from Transformers (BERT), augmented with adversarial training via the Fast Gradient Sign Method (FGSM), aiming to enhance scoring accuracy and stability in multi-dimensional AES tasks. Findings reveal that *BERT-Integrated-MTL-FGSM* significantly outperforms existing baselines. Integrating Explainable Artificial Intelligence (XAI) elucidates the complex evaluation processes within the deep learning black box, offering potentially diagnostic feedback for students and assisting educators in delivering individualized instruction guided by word-level attributions across diverse analytic rating scales. These insights enhance the robustness, generalizability, and interpretability of multi-dimensional AES models, thus paving an interdisciplinary pathway for multi-dimensional writing assessment and instruction.

Keywords: adversarial training; automated essay scoring; explainable artificial intelligence; multi-dimensional writing assessment; multi-task learning

1. Introduction

Writing is a pivotal component of educational instruction and assessment, recognized for its role as an essential linguistic skill that encompasses complex cognitive functions such as critical reasoning, logical thinking, and problem-solving (Sari, Utomo and Astina 2021; Uto 2021). The integration of writing tests into large-scale assessment programs has spurred the advancement of automated essay scoring (AES) systems, offering a practical alternative to the laborious and time-consuming process of manual grading.

Over the past decades, AES has garnered interdisciplinary interest due to its substantial commercial and educational applications (Latifi and Gierl 2021; Shermis 2014). From the perspective of rating traits, numerous studies have focused on constructing AES models based on single holistic scores provided by human raters (Goh, Sun and Yang 2020; Vajjala 2018). However, studies that delve into AES models incorporating analytic rating traits remain relatively scarce (Kumar, Fraser and Boulanger 2017; Taghipour 2017).

© The Author(s), 2026. Published by Cambridge University Press. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

Downloaded from https://www.cambridge.org/core. 19 Nov 2026 at 19:13:46, subject to the Cambridge Core terms of use.

Regarding approaches to AES, two types can be categorized: traditional approaches with manually crafted linguistic features and neural network approaches without feature engineering. Traditional feature-based approaches employ multiple regression models to predict essay scores (Lee, Ge and Chung 2021; MacArthur, Jennings and Philippakos 2019). Despite significant advancements, traditional approaches require considerable effort in model construction and feature selection when applied to different genres or rating traits. This necessitates starting anew to select the appropriate linguistic features to fit the specific model.

By eliminating the need for manual feature selection, neural network approaches streamline the modeling process and enhance the scalability and adaptability of AES systems across various genres and rating scales (Beseiso, Alzubi and Rashaideh 2021; Kumar and Boulanger 2021). However, most neural network approaches primarily focus on holistic scoring, often overlooking the benefits of multi-task learning (MTL) and the application of pre-trained language models in multi-dimensional AES tasks (Blanchard *et al.* 2013; Vajjala and Lučić 2018). This gap underscores the need for a deeper understanding of different constructs of writing quality, potentially leading to more nuanced and effective AES implementations. Besides, despite the broad applicability of neural networks, their vulnerability to adversarial examples of input perturbations causing errors poses a significant challenge, particularly in sensitive applications like AES.

To bridge the identified gaps, rather than proposing novel individual components in isolation, this study builds a unified framework whose methodological originality lies in the targeted integration of MTL, adversarial robustness training, and word-level explainability on top of the pre-trained language model, Bidirectional Encoder Representations from Transformers (BERT). While designed for multi-dimensional AES, this integrated framework also serves as a methodological reference for broader natural language processing (NLP) tasks that require multifaceted evaluation coupled with robustness and post-hoc interpretability. The main contributions of this study can be summarized as follows:

- **Unified multi-trait modeling.** We propose an integrated MTL approach to fine-tune BERT across both holistic and analytic rating scales. While joint modeling of holistic and analytic scores has been explored previously, our formulation explicitly derives the holistic score as the arithmetic sum of the four analytic scores, leveraging shared representations to improve generalizability across diverse writing constructs without manual feature engineering.
- **Adversarial robustness for evaluative stability.** We address the vulnerability of AES models to subtle input manipulations by embedding Fast Gradient Sign Method (FGSM) adversarial training within the multi-task framework, which remains underexplored in AES, thereby maintaining the integrity and robustness of the scoring process.
- **Granular attribution for pedagogical utility.** We augment the framework with explainable artificial intelligence (XAI) via Integrated Gradients (IG), attributing multi-dimensional scoring decisions to individual input tokens and providing potentially diagnostic and tailored feedback for multi-dimensional writing assessment and instruction.

We emphasize that the individual building blocks employed here—MTL, FGSM-based adversarial training, IG-based attribution, and BERT fine-tuning—are each well established in the broader NLP literature. The originality of this study, therefore, lies not in any single component, but in their targeted integration: to the best of our knowledge, this is the first AES framework that simultaneously couples joint holistic–analytic scoring with adversarial robustness training and word-level attribution. In particular, while adversarial training is mature in general NLP, its application to multi-dimensional essay scoring has remained largely unexplored. Section 3.5 further provides a conceptual account of why these components reinforce one another when built upon

Downloaded from https://www.cambridge.org/core. 19 Nov 2026 at 19:13:46, subject to the Cambridge Core terms of use.

a shared differentiable encoder. This integration yields empirical and diagnostic insights across analytic rating scales that none of the components affords in isolation.

2. Related work

2.1 Rating scales for writing evaluation

A well-defined set of rating scales is essential for the validity of essay evaluation. Scholars have emphasized its role in assisting raters and educators in generating consistent scores and minimizing biases stemming from subjective judgments (Wind 2020). Holistic scoring, which assigns a single score to reflect overall writing quality, is valued for its efficiency and cost-effectiveness, especially in large-scale educational assessments. This method simplifies the scoring process by allowing raters to assign a single score quickly without needing multiple readings to analyze diverse analytical scales and descriptors. However, it faces substantial criticism for its lack of granularity. Specifically, holistic scoring often fails to provide detailed feedback for students, especially those struggling with writing, regarding specific areas for improvement (Weigle 2002). Moreover, it does not allow evaluators to recognize and value the unique characteristics that differentiate essays.

Scholars today recognize the significant value of analytic rating scales in writing assessments, particularly for their capacity to deliver detailed and diagnostic feedback across various dimensions of writing quality (Zou, Yan and Fan 2024). Analytic scoring excels in providing a nuanced and comprehensive portrayal of student performance by considering multiple aspects of writing quality. It not only facilitates pinpointing specific developmental needs but also bolsters formative assessment practices by addressing individual performance aspects distinctly and enabling personalized feedback (Wind 2020). Therefore, the present study combines the advantages of holistic and analytic scoring methods to enhance reliability and validity.

2.2 Approaches to automated essay scoring

Traditional feature-based models emphasized the correlation between linguistic features and the overall quality of writing. (Crossley 2020; Latifi and Gierl 2021), they have yet to be explored in predicting specific writing quality constructs. Besides, their generalizability is often constrained by the variability of linguistic features across different genres and rating scales, requiring significant efforts to adapt and refine these systems for broader applications.

Recent advancements in AES have markedly enhanced model generalizability across diverse genres and topics, primarily through neural network approaches. Unlike traditional AES methods that rely heavily on handcrafted linguistic features, neural networks leverage deep learning to automatically extract and learn from raw text data, thus bypassing the need for predefined feature sets. Scholars have adopted the convolutional neural network approach (Dong and Zhang 2016), the recurrent neural network approach (Park, Lee and Shin 2021), and the deep convolution recurrent neural network approach (Dasgupta *et al.* 2018) to explore AES tasks.

While the shift toward increasingly sophisticated neural models promises more reliable and insightful assessments of student writing, it concurrently introduces inherent interpretability challenges—often characterized as the black box problem. This opacity motivates the explainability methods reviewed in Section 2.6. Furthermore, although hybrid approaches have garnered significant interest for their broader potential in NLP tasks (Dasgupta *et al.* 2018; Uto, Xie and Ueno 2020), there remains a pressing need to investigate their efficacy in multi-dimensional essay scoring across diverse genres and rating traits.

2.3 Pre-trained language models in AES tasks

Although neural network architectures have achieved remarkable success in AES tasks (Li, Chen and Nie 2020; Uto 2021), their heavy reliance on extensively annotated datasets remains a

Downloaded from https://www.cambridge.org/core. 19 Nov 2026 at 19:13:46, subject to the Cambridge Core terms of use.

significant bottleneck. To address this limitation, scholars have demonstrated that pre-trained language models can be fine-tuned for different NLP tasks (Devlin *et al.* 2018; Sung *et al.* 2019). BERT, introduced by Google AI Language, has set new performance benchmarks across multiple tasks and has proven helpful in AES (Mayfield and Black 2020; Nadeem *et al.* 2019) and automated short-answer grading (Lun *et al.* 2020; Sung *et al.* 2019).

MTL offers a promising avenue to exploit such pre-trained models further, as shared representations across related tasks improve both generalization and training efficiency (Bernardini *et al.* 2021; Gao *et al.* 2023). Building upon these advancements, recent literature has explored BERT-based architectures for the joint modeling of holistic and analytic essay scores via multi-task or multi-loss learning paradigms. For instance, Wang *et al.* (2022) optimized multi-scale essay representations using diverse loss objectives; Kumar *et al.* (2022) framed trait-specific scoring as auxiliary tasks to enhance holistic prediction; and Cho *et al.* (2024) introduced a dual-scale BERT incorporating multi-trait representations for concurrent holistic and dimension-specific grading.

While these foundational studies establish the viability of joint holistic–analytic modeling using pre-trained encoders, they leave notable structural and interpretability gaps unexplored. Specifically, they neither enforce a constraint in which the holistic score is compositionally derived from analytic dimensions, nor do they couple joint scoring with adversarial robustness training and gradient-based attribution.

2.4 Large language models and ensemble approaches in AES

Large language models (LLMs) and ensemble methodologies are rapidly reshaping automated assessment. Contemporary studies have explored large-scale generative models, such as GPT-3.5 and GPT-4, in zero-shot and prompt-based essay evaluation (Xia, Mao and Zheng 2024), with comparative evidence suggesting that generative models can surpass traditional neural baselines in specific settings (Kim and Jo 2024; Li and Liu 2024). Systematic assessments have further investigated their accuracy, validity, and fairness across diverse scoring rubrics (García-Varela *et al.* 2025; Huang and Wilson 2025; Pack, Barrett and Escalante 2024), while highlighting the critical role of prompt design for open-source LLMs (Abujadallah, Saad and Abudalfa 2025). Beyond the immediate context of educational assessment, LLMs and advanced representation learning architectures have demonstrated substantial predictive utility across adjacent domains. For instance, transfer learning from pre-trained models has proven effective for financial sentiment analysis (Ergun and Sefer 2025), while tailored data representations have markedly enhanced financial statement fraud detection (Alaygut and Sefer 2025). These interdisciplinary successes underscore the cross-domain versatility of the representation-learning paradigm upon which our proposed framework is built. Concurrently, ensemble approaches combining BERT, XGBoost, and neural networks with ridge regression have been shown to substantially reduce root mean squared error (RMSE) in holistic scoring (Poonpon and Chansanam 2025).

Although these generative and ensemble methods define the current frontier, they often incur substantial computational overhead, depend heavily on prompt engineering, and offer limited access to gradient-based token-level attribution. In contrast, the present study positions a compact, fully differentiable BERT-based framework as an efficient, controllable, and interpretable alternative, prioritizing word-level attribution and computational economy for robust real-world deployment.

2.5 Adversarial training

Despite their advanced capabilities, deep neural networks are particularly vulnerable to adversarial examples—inputs that have been subtly and deliberately manipulated to deceive models during testing (Shrivastava *et al.* 2017; Wong, Rice and Kolter 2020). These adversarial examples are engineered by introducing minute perturbations to the input data of models, specifically designed to maximize the loss and induce misclassification (Goodfellow, Shlens and Szegedy 2014). Even the

Downloaded from https://www.cambridge.org/core. 19 Nov 2026 at 19:13:46, subject to the Cambridge Core terms of use.

most advanced algorithms frequently struggle to accurately identify these modified inputs, which typically remain undetectable to the human eye.

Adversarial training contributes to the robustness of multi-dimensional AES models by equipping them to handle various types of manipulative inputs. To mitigate these vulnerabilities, the present study integrates adversarial training into the multi-dimensional AES pipeline, aiming to develop robust models capable of withstanding input perturbations without compromising scoring accuracy or stability.

2.6 Interpretability of AES models

The inherently opaque nature of deep neural networks has spurred the development of various explanation methods, highlighting a growing need to demystify the decision processes of these complex models (Kolek *et al.* 2022). XAI has become a crucial field within AI-driven educational scenarios, pioneering innovative methodologies to clarify and interpret the predictions of complex models. Advances in XAI enhance transparency and deepen understanding of the decision-making mechanisms, contributing to more accountable and comprehensible technological solutions (Holzinger *et al.* 2022).

Building trust remains a significant challenge, as the AES black box often remains opaque to educators and students. It underscores the need for continued research in XAI to make AES models more interpretable and aligned with educational goals, thereby enhancing their utility and acceptance in educational settings. A further challenge lies in elucidating the complete decision-making process within AES pipelines, spanning from holistic to analytic scoring. By integrating XAI into these frameworks, stakeholders—including educators and students—can gain insights into the underlying rationale driving specific scoring decisions.

Taken together, these intersecting research strands motivate the integrated computational framework developed in the subsequent sections.

3. Method

3.1 Dataset

The Automated Student Assessment Prize (ASAP) dataset, which includes a diverse array of essay topics and genres, has played a pivotal role in both benchmarking and advancing AES models. It stands as a cornerstone for developing and validating machine learning algorithms. It comprises essays across eight essay prompts authored by students in grades seven through ten. Specifically, four of these prompts are source-based, requiring students to refer to a document provided within the question. In contrast, the remaining four prompts encompass traditional genres such as persuasive, expository, and narrative essays. These collections exhibit a broad spectrum of subjects and writing modalities. The essay lengths in this dataset exhibit significant variability, with averages ranging from 150 to 550 words per response.

Dataset 7 was specifically chosen for its comprehensive scoring rubric. This scoring was executed across four essential dimensions: *Ideas*, *Organization*, *Style*, and *Conventions*, with scores for each trait ranging from 0 to 3. It comprises 1,569 narrative essays written by seventh-grade students on the topic of *Patience* (Shin and Gierl 2021). These essays are characterized by both brevity and depth, averaging approximately 250 words in length. The scoring distribution across each rating scale is illustrated in Figure 1. Results reveal that a substantial number of students fall within the intermediate category in terms of writing quality.

3.2 Integrated multi-task learning with fine-tuning BERT

The present study employed an integrated MTL framework (Figure 2), supported by fine-tuning the BERT model, to simultaneously predict scores across different constructs of writing quality.

Downloaded from https://www.cambridge.org/core. 19 Nov 2026 at 19:13:46, subject to the Cambridge Core terms of use.

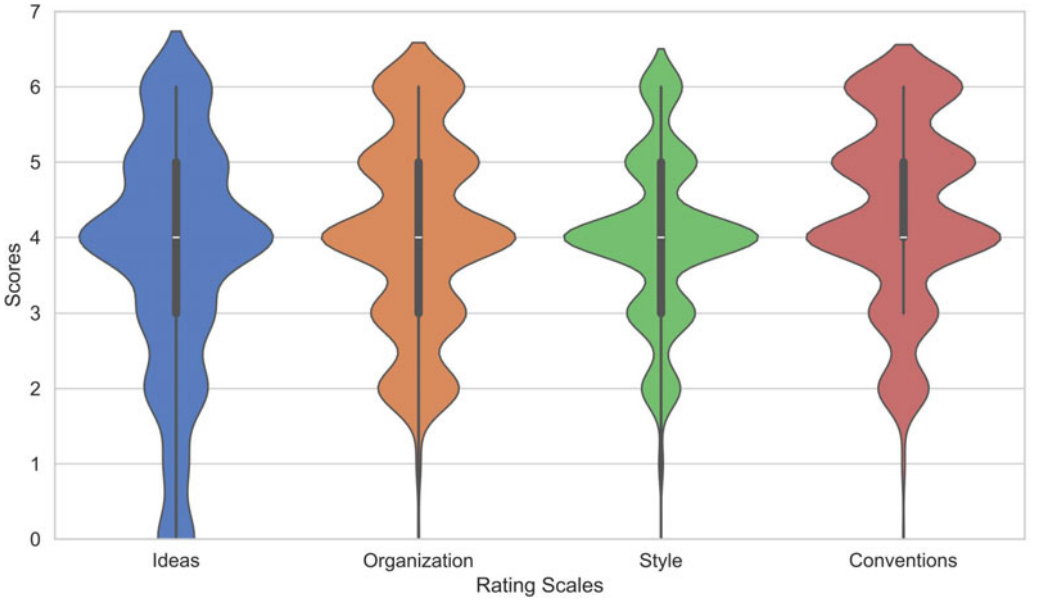


Figure 1. Score distribution for different rating scales of ASAP dataset 7.

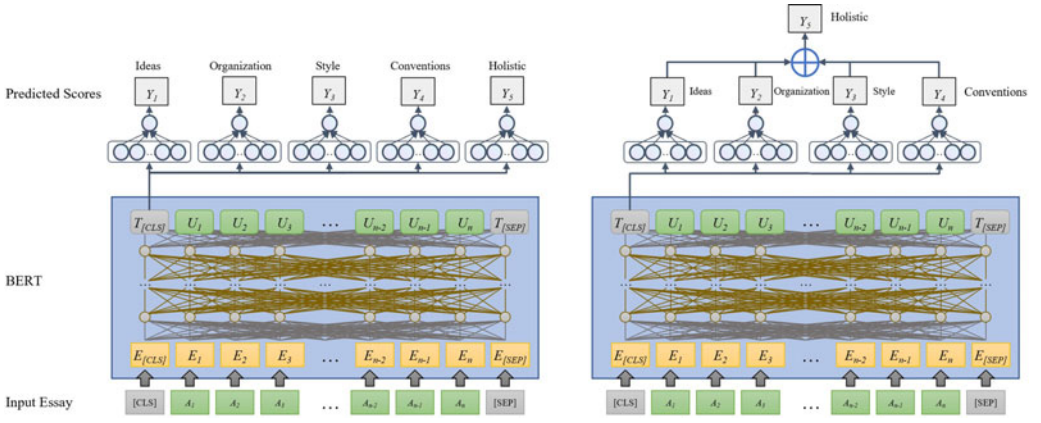


Figure 2. Integrated MTL with fine-tuning BERT for multi-dimensional AES (Left: *BERT-Simple-MTL*; Right: *BERT-Integrated-MTL*).

We propose the *BERT-Simple-MTL* approach, which treats each rating scale (including *Ideas*, *Organization*, *Style*, *Conventions*, and *Holistic*) as an independent yet simultaneous task. We further propose *BERT-Integrated-MTL*, which explores the interrelationships between holistic and analytic rating scales. Within this model, the holistic score is computed as the cumulative sum of the analytic scores for *Ideas*, *Organization*, *Style*, and *Conventions*.

By incorporating MTL with fine-tuning BERT, the present study benefits multi-scale AES tasks through the shared representation of each scale compared with single-scale scoring tasks. To model the intricate relationship between the input essay, denoted as A, and its evaluation metrics, we employ the BERT algorithm to systematically extract salient attributes from A. These attributes are subsequently employed to predict ratings across multiple dimensions. The mathematical

Downloaded from https://www.cambridge.org/core. 19 Nov 2026 at 19:13:46, subject to the Cambridge Core terms of use.

framework governing this process is encapsulated as follows:

$$\begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \\ Y_5 \end{pmatrix} = f_{\text{BERT-Integrated-MTL}}(A), \quad \text{where} \quad Y_5 = Y_1 + Y_2 + Y_3 + Y_4,$$

and $f_{\text{BERT-Integrated-MTL}}(A)$ represents the integration of the BERT framework with the input essay A, yielding a set of attributes. These attributes are then exploited to calculate scores Y_1, Y_2, Y_3, Y_4 corresponding to the different constructs of writing quality regarding *Ideas*, *Organization*, *Style*, and *Conventions*. The aggregate score Y_5 is derived by summing these individual scores, amalgamating the discrete evaluations into an overarching assessment criterion.

Given the categorical nature of the proposed AES model, we employ categorical cross-entropy as the loss function for each dimension, a standard method widely employed in multi-class categorization tasks. The aggregate loss L for the entire framework is computed by summing the losses for each rating scale and the holistic score:

$$L = \sum_{i=1}^{4} L_i(Y_i, \hat{Y}_i) + L_5(Y_5, \hat{Y}_5),$$

capturing the cumulative impact of discrepancies across all rating scales, including *Ideas*, *Organization*, *Style*, and *Conventions*, as well as the holistic score Y_5.

Each L_i denotes the loss for dimension i, where Y_i represents the true score assigned by human raters and $\hat{Y}_i$ denotes the predicted scores by the proposed AES model. The loss functions L_i are delineated as:

$$L_i = -\sum_{c=1}^{C} p_{ic} \log q_{ic},$$

here, p_{ic} represents the actual probability of class c within dimension i, and q_{ic} signifies the predicted probability of class c.

3.3 Adversarial training: the Fast Gradient Sign Method

Adversarial training emerges as a pivotal approach in enhancing the robustness of machine learning models against input perturbations. FGSM is an advanced technique for generating adversarial examples—inputs deliberately designed to disrupt machine learning model outputs (Goodfellow *et al*. 2014). During the training phase, gradients of the loss function in relation to the input essay text are carefully calculated to discern how minor textual changes can substantially affect scoring outcomes across all rating scales. Utilizing these gradients, adversarial examples are generated by introducing perturbations scaled by ϵ. By incorporating these adversarial examples, the model is exposed to both original and strategically modified versions of the essays.

The fundamental principle of FGSM is to generate adversarial examples by applying minimal yet strategically optimized perturbations to the input data. The mathematical formulation of FGSM is expressed as follows:

$$\mathbf{x}^{adv} = \mathbf{x} + \epsilon \cdot \text{sign}(\nabla_{\mathbf{x}} J(\theta, \mathbf{x}, y)),$$

where $\mathbf{x}^{adv}$ signifies the adversarial example derived from the original input $\mathbf{x}$, ϵ a small scalar that determines the perturbation's magnitude, $\nabla_{\mathbf{x}} J(\theta, \mathbf{x}, y)$ denotes the gradient of the loss function J

Downloaded from https://www.cambridge.org/core. 19 Nov 2026 at 19:13:46, subject to the Cambridge Core terms of use.

with respect to the input $\mathbf{x}$, θ represents the model's parameters, and y is the true label for $\mathbf{x}$. The sign() function directs the perturbation towards increasing the loss, thus testing the model's robustness in adversarial settings.

It is worth clarifying our rationale for adopting FGSM. We deploy FGSM not as the most potent adversarial attack available, but rather as a computationally lightweight, single-step perturbation scheme. First, FGSM injects adversarial noise utilizing a single gradient computation, which preserves the computational tractability of the multi-task training configuration (see Table 7). Second, its algorithmic simplicity facilitates the isolation and reproducibility of robustness gains, thereby establishing a controlled baseline for embedding-space perturbations in AES. Third, while the core attack mechanism is standard, the perturbations are computed against the joint multi-task loss; consequently, each adversarial example simultaneously challenges all five scoring dimensions through the shared encoder (see Section 3.5). Multi-step alternatives, such as Projected Gradient Descent (PGD) and FreeLB, explore a more comprehensive adversarial neighborhood at substantially higher computational cost; we evaluate both empirically in Section 4.5. Thus, we position FGSM as an efficient, foundational entry point for adversarial training in AES, rather than as an upper bound on defensive capability.

3.4 Interpretability via integrated gradients

In educational assessment, the transparency and interpretability of AES models are essential for establishing trust and ensuring accountability. Interpretability enables stakeholders to verify the model's underlying logic, confirm its alignment with pedagogical objectives, and identify potential biases or spurious correlations in its reasoning. In this study, we leverage IG (Ross and Doshi-Velez 2018; Sundararajan, Taly and Yan 2017) to generate word-level attributions, aiming to support students' writing development through data-driven feedback while providing educators with insight into the model's scoring behavior. We adopt IG as an applied, post-hoc interpretability technique rather than proposing a methodologically novel XAI algorithm. Furthermore, it is important to note that IG provides local, instance-specific feature attributions rather than global decision transparency; the resulting attributions highlight how individual tokens contribute to the prediction for a specific essay, rather than extracting universal causal rules across the entire corpus. Nevertheless, this local interpretability helps bridge the gap between highly performant but opaque neural architectures and the pedagogical need for granular, potentially diagnostic feedback. This dual advantage of automated scoring and token-level attribution enhances the educational utility of the proposed framework, positioning it as a promising tool for both large-scale assessment and personalized instruction.

3.5 Theoretical rationale for the integrated framework

While the constituent techniques of our framework are individually well established, their integration is not an ad hoc combination. Instead, all three components converge upon and mutually reinforce a single computational quantity: the gradient of a shared differentiable encoder. First, the integrated MTL architecture constrains the holistic score to be the cumulative sum of the four analytic scores. This structural constraint serves as an explicit inductive bias. It compels the shared encoder to learn representations where analytic dimensions compose linearly to reflect overall writing quality, thereby regularizing all five scoring tasks toward a unified construct and mitigating task-specific overfitting (Zhang and Yang 2021).

Second, FGSM adversarial training augments the optimization objective with perturbations derived from the sign of the input gradient. To a first-order approximation, minimizing this adversarial loss implicitly penalizes the gradient norm, promoting local smoothness of the scoring function within the embedding space (Goodfellow *et al.* 2014; Miyato, Dai and Goodfellow 2016). Crucially, this smoothness constraint operates on the shared representation underlying all

Downloaded from https://www.cambridge.org/core. 19 Nov 2026 at 19:13:46, subject to the Cambridge Core terms of use.

five tasks, which facilitates the transfer of robustness gains across rating dimensions rather than isolating them within a single task.

Finally, IG leverages this exact gradient pathway for post-hoc attribution (Sundararajan *et al*. 2017). The smoother loss surface cultivated by adversarial training yields more stable input gradients, which in turn produce more reliable word-level attributions for instructional feedback. In summary, MTL supplies the cross-dimensional inductive bias, adversarial training enforces the local stability of the shared representation, and IG extracts these stabilized gradients for interpretable evaluation. Consequently, the three components form a mutually supportive architecture in which each element fortifies the theoretical preconditions of the others, rather than a mere concatenation of disparate methods.

3.6 Evaluation metrics

To gauge the alignment between the scores predicted by the computational model and those graded by human evaluators, we employed four distinct metrics to provide a robust framework for evaluating the precision and reliability of the model's performance in mirroring human judgment.

R^2 was a goodness-of-fit measure to evaluate the proportion of variance in human raters' scores that the model can explain. Its formula is given by:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2},$$

where y_i represents the actual scores assigned by human raters, serving as the benchmark for evaluating the AES model's accuracy. The model aims to closely align its predicted scores $\hat{Y}_i$ with these actual scores to effectively mimic human scoring behavior. $\bar{y}$ denotes the average of all actual scores, offering a reference point for assessing variations in both actual and predicted scores.

Quadratic Weighted Kappa (QWK) was employed to measure the degree of agreement between the evaluations of the two raters.

$$QWK = 1 - \frac{2 \times \sum_{i,j} w_{ij} o_{ij}}{\sum_{i,j} w_{ij} e_{ij}},$$

where w_{ij} denotes the weight assigned to the disagreement between raters for terms categorized i and j, o_{ij} represents the observed frequency of agreement, and e_{ij} signifies the expected frequency of agreement under the assumption of chance.

Exact Agreement is a metric designed to quantify the proportion of essays where the scores assigned by human raters perfectly align with those predicted by the model.

Exact-Plus-Adjacent Agreement broadens the criteria by measuring the percentage of essays that either receive the same score or within one point by both the human raters and the model.

3.7 Hyperparameter selection and experimental setup

All experiments were implemented with the BERT-base-uncased encoder (12 layers, 768 hidden dimensions, 110 M parameters). Models were trained with the AdamW optimizer for 5 epochs with a batch size of 32 and a maximum sequence length of 128 tokens, on a single NVIDIA RTX 3090 GPU (24 GB). The learning rate schedule used a linear warmup over the first 10% of training steps followed by linear decay, a weight decay of 0.01, and a dropout rate of 0.1.

Two hyperparameters proved most consequential and were selected via grid search under the same clean-test evaluation condition as the main results: the FGSM training perturbation radius ϵ and the learning rate. ϵ here denotes the perturbation radius applied during adversarial *training*, as distinct from the test-time perturbation analysis reported in Section 4. As shown in Table 1, ϵ was examined over {0.00, 0.05, 0.10, 0.15, 0.20, 0.25, 0.30}: holistic performance peaks

Downloaded from https://www.cambridge.org/core. 19 Nov 2026 at 19:13:46, subject to the Cambridge Core terms of use.

Table 1. Hyperparameter sensitivity of BERT-Integrated-MTL-FGSM on the holistic dimension. Selected values are shown in bold

Parameter	Value	R^2	QWK	Exact	Exact + adj.
FGSM ϵ	0.00	0.820	0.890	0.280	0.670
FGSM ϵ	0.05	0.840	0.900	0.320	0.690
FGSM ϵ	**0.10**	0.849	0.910	0.338	0.710
FGSM ϵ	0.15	0.850	0.880	0.300	0.710
FGSM ϵ	0.20	0.810	0.840	0.280	0.680
FGSM ϵ	0.25	0.780	0.790	0.170	0.610
FGSM ϵ	0.30	0.620	0.680	0.130	0.560
Learning rate	1e-5	0.832	0.899	0.324	0.692
Learning rate	**2e-5**	0.849	0.910	0.338	0.710
Learning rate	3e-5	0.842	0.903	0.331	0.701
Learning rate	5e-5	0.821	0.881	0.309	0.670

at $\epsilon = 0.10$ (QWK $= 0.910$) and degrades sharply beyond $\epsilon = 0.15$, falling to a QWK of 0.680 at $\epsilon = 0.30$, indicating that excessive perturbations distort the semantic content of the embedding space rather than regularizing it (Goodfellow *et al.* 2014). The learning rate was examined over {1e−5, 2e−5, 3e−5, 5e−5}, with 2e−5 producing the strongest holistic QWK—consistent with the fine-tuning ranges recommended for BERT (Devlin *et al.* 2018). Accordingly, the final BERT-Integrated-MTL-FGSM configuration adopts $\epsilon = 0.10$ and a learning rate of 2e−5.

4. Results and discussion

4.1 Analysis of integrated multi-task learning and adversarial training with fine-tuning BERT for AES tasks

To validate the effectiveness of our proposed AES model for multi-dimensional writing assessment, we conducted a thorough comparative study to evaluate the effectiveness of MTL with fine-tuning BERT for multi-scale tasks. The present study involved three distinct configurations as follows: *BERT-Simple-MTL*, *BERT-Integrated-MTL* and *BERT-Integrated-MTL-FGSM*. Table 2 indicates the effectiveness of different approaches to multi-dimensional writing assessment.

Results demonstrate that *BERT-Simple-MTL* demonstrates moderate to high performance across different rating scales. It excels at the *Holistic* scale with an R^2 of 0.791 and a QWK score of 0.873. The higher scores in Exact-Plus-Adjacent Agreement across all rating scales indicate a general alignment with human raters, although not always precise. Potential reasons for the observed limitations in achieving high Exact Agreement for holistic scoring may stem from the independent scoring mechanism utilized by the model. Additionally, the broad scoring range, which extends from 0 to 24 points, likely complicates the precise evaluation of essays, as it demands a nuanced understanding and integration of multiple assessment dimensions into a single aggregate score.

BERT-Integrated-MTL-FGSM, which incorporates adversarial training, exhibits superior performance across nearly all rating scales according to various evaluation metrics. Notably, it achieves the highest R^2 and QWK in the *Holistic* dimension (0.849 & 0.910 respectively). The improvements in both Exact Agreement and Exact-Plus-Adjacent Agreement across different rating scales underscore the substantial benefits of adversarial training. It substantially

Downloaded from https://www.cambridge.org/core. 19 Nov 2026 at 19:13:46, subject to the Cambridge Core terms of use.

Table 2. Comparison of different approaches to multi-scale scoring tasks (E: Exact Agreement; E + A: Exact-Plus-Adjacent Agreement)

Approaches	Rating scales	R^2	QWK	E	E + A
BERT-simple-MTL	Ideas	0.790	0.862	0.643	0.955
	Organization	0.701	0.800	0.672	0.959
	Style	0.755	0.854	0.752	0.987
	Conventions	0.733	0.822	0.675	0.962
	Holistic	0.791	0.873	0.283	0.662
BERT-integrated-MTL	Ideas	0.799	0.861	0.662	0.939
	Organization	0.739	0.820	0.675	0.971
	Style	0.765	0.856	0.758	0.987
	Conventions	0.756	0.817	0.653	0.968
	Holistic	0.807	0.884	0.315	0.659
BERT-integrated-MTL-FGSM	Ideas	0.819	0.884	0.697	0.959
	Organization	0.766	0.834	0.691	0.975
	Style	0.816	0.890	0.799	0.994
	Conventions	0.801	0.873	0.761	0.971
	Holistic	0.849	0.910	0.338	0.710

enhances the model's resilience against perturbations, thereby increasing the precision and consistency of multi-dimensional scoring. Moreover, it promotes scoring stability across broad response groups (see Section 4.6), an essential characteristic for navigating the complex and varied nature of multi-dimensional writing assessment.

We have visualized the confusion matrices comparing true and predicted scores for *BERT-Integrated-MTL-FGSM*, as depicted in Figure 3. These matrices demonstrate that the predictions from *BERT-Integrated-MTL-FGSM* align more closely with the actual scores, supporting our earlier observations. This evidence underscores the superior performance of *BERT-Integrated-MTL-FGSM*, affirming the value of integrated MTL with fine-tuning BERT and adversarial training in multi-dimensional scoring tasks.

To gain deeper insights into the model's scoring behavior, we analyzed instances of inaccurate predictions. This analysis reveals two systematic divergences from human evaluative behavior. First, the model exhibits systematic *stringency*, adopting a stricter scoring regimen than human evaluators: it tends to assign the lowest possible scores to essays that do not meet specified word count requirements, reflecting a rigid adherence to predefined criteria, and it penalizes spelling and word formation errors more heavily than human raters typically would. Second, the model is prone to *score compression*, an inclination toward middle-level scores. From a machine learning optimization perspective, predicting middle-level scores minimizes the overall expected deviation from true scores across the dataset. While this trade-off optimizes aggregate performance metrics (e.g., reducing the mean absolute error (MAE)), it may sacrifice absolute precision at the extreme ends of the scoring spectrum for individual outlier essays. These qualitative observations are quantitatively corroborated by the stratified error analysis in Section 4.6, where middle-band essays exhibit the lowest absolute errors.

Downloaded from https://www.cambridge.org/core. 19 Nov 2026 at 19:13:46, subject to the Cambridge Core terms of use.

Table 3. Comparison of QWK scores at the holistic rating scale with *BERT-Integrated-MTL-FGSM* and other well-established baselines

Approaches	QWK
CNN-LSTM-Att (Dong *et al.* 2017)	0.801
TSLF-ALL (Liu *et al.* 2019)	0.801
Qe-C-LSTM (Dasgupta *et al.* 2018)	0.815
YAKE-T BERT (Wu *et al.* 2022)	0.837
Tran-BERT-MS-ML-R (Wang *et al.* 2022)	0.839
BERT-integrated-MTL-FGSM	**0.910**

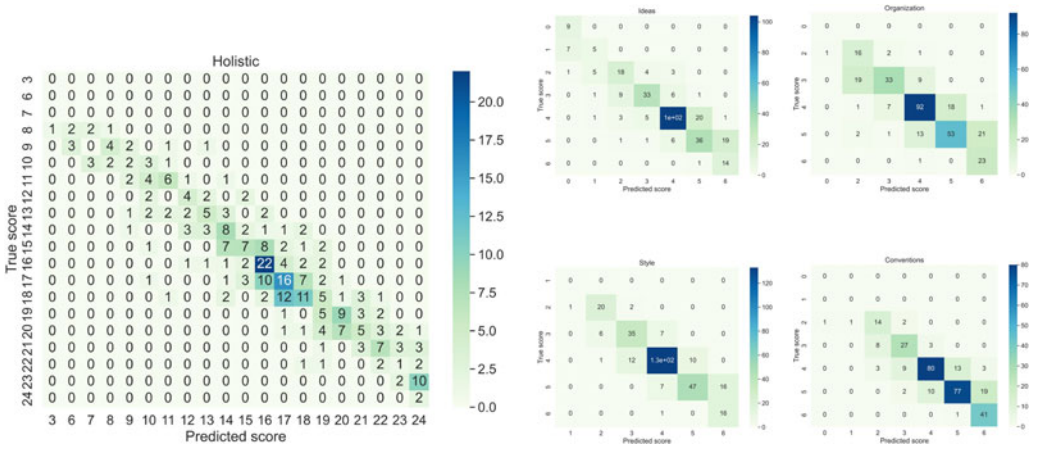


Figure 3. Confusion matrices for *BERT-Integrated-MTL-FGSM* across multi-dimensional writing assessment.

We also compared the results with other well-established baseline models. Scholars have employed neural network approaches to develop AES models, contributing to a rich landscape of approaches in this field. Table 3 demonstrates the QWK scores for the holistic rating scale with *BERT-Integrated-MTL-FGSM* and other well-established baselines. CNN-LSTM-Att (Dong, Zhang and Yang 2017) integrates CNN and LSTM layers with an attention mechanism to specifically address AES tasks, demonstrating the utility of hierarchical network structures. TSLF-ALL (Liu, Xu and Zhu 2019) explores a dual-phase learning architecture that capitalizes on the benefits of feature-engineered methods to enhance interpretability and performance. Qe-C-LSTM (Dasgupta *et al.* 2018) employs a sophisticated deep convolutional recurrent neural network to achieve greater accuracy in essay scoring. YAKE-T BERT (Wu *et al.* 2022) is fine-tuned exclusively on student essays, eschewing any contextual information from essay prompts to concentrate solely on the evaluation of text. Tran-BERT-MS-ML-R (Wang *et al.* 2022) employs two BERT models to derive features at the token, document, and segment levels, utilizing multiple loss functions to score essays effectively.

Results revealed that *BERT-Integrated-MTL-FGSM* achieved a QWK score of 0.910, surpassing the benchmark established by Wang *et al.* (2022) with a QWK score of 0.839. A paired t-test analysis confirmed that *BERT-Integrated-MTL-FGSM* significantly improved performance compared with the most robust baseline ($p < 0.0010$). This achievement signifies a noteworthy improvement in holistically predicting essay quality, marking a 7.1% enhancement over the most proficient

Downloaded from https://www.cambridge.org/core. 19 Nov 2026 at 19:13:46, subject to the Cambridge Core terms of use.

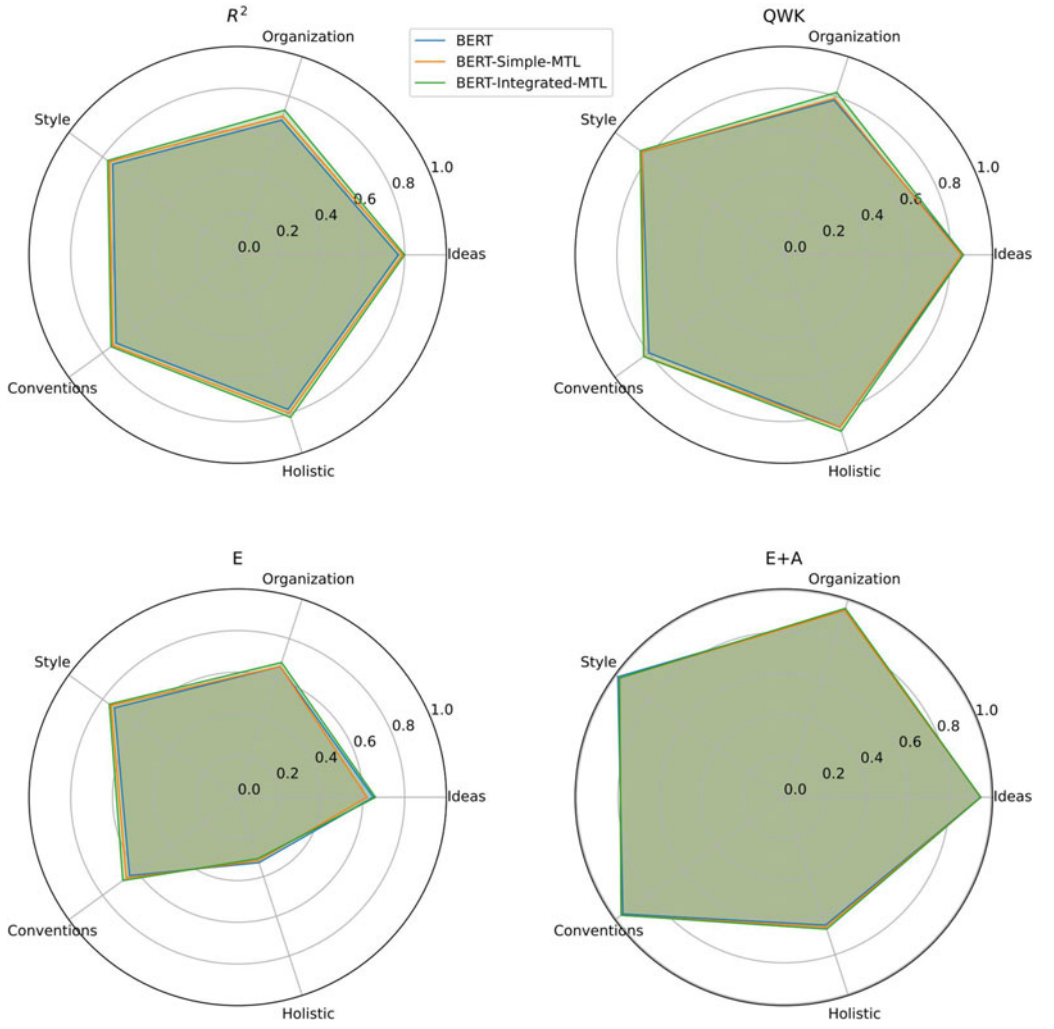


Figure 4. Radar charts for different approaches across multi-scale scoring tasks.

baseline model. Results also underscore the efficacy of integrating MTL and adversarial training techniques in enhancing the precision and reliability of AES models.

4.2 Comparison of effectiveness of multi-task learning in multi-dimensional writing assessment

To explore the effectiveness of MTL in multi-dimensional writing assessments, we compare the performance of three different approaches regarding *BERT*, *BERT-Simple-MTL*, and *BERT-Integrated-MTL*. *BERT* here refers to the specific adaptation of the BERT model for AES tasks, where it is fine-tuned independently for each individual rating scale. To uncover data patterns and correlations, we utilized radar charts to visualize multivariate data, facilitating a comparative analysis of three different BERT model adaptations: *BERT*, *BERT-Simple-MTL*, and *BERT-Integrated-MTL* across various analytic rating scales (see Figure 4).

BERT-Integrated-MTL consistently outperforms *BERT* and *BERT-Simple-MTL* across diverse evaluation metrics and rating scales in AES tasks. Specifically, *BERT* generally displays the smallest area coverage, particularly in the Holistic and Conventions dimensions, which implies

Downloaded from https://www.cambridge.org/core. 19 Nov 2026 at 19:13:46, subject to the Cambridge Core terms of use.

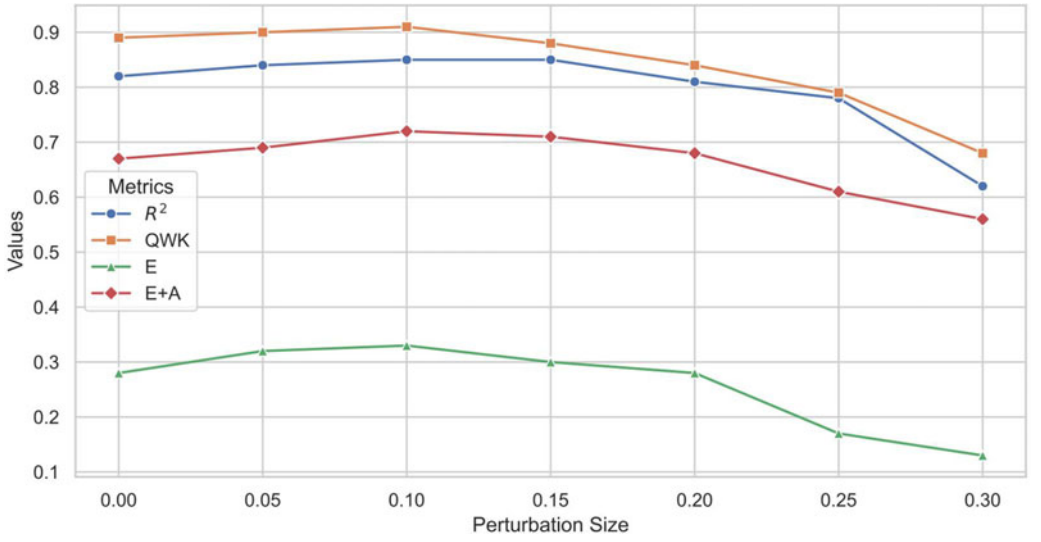


Figure 5. Performance of *BERT-Integrated-MTL-FGSM* across various perturbation sizes at the holistic scale.

reduced reliability and lower concordance with human assessments. The superior performance of *BERT-Integrated-MTL* can be attributed to its integration of MTL capabilities, which adeptly incorporates various constructs of writing quality into the model's framework. The enhanced performance is likely due to its capacity to capture complex interdependencies among scoring criteria, thereby improving its generalizability. Moreover, the increased accuracy observed in *BERT-Integrated-MTL* may stem from its ability to leverage shared representations across multi-scale scoring tasks. These insights highlight the significance of integrating advanced computational techniques in AES systems to more accurately replicate the nuanced judgment of human evaluators, which is crucial for developing more precise and dependable AES models.

4.3 Comparison of effectiveness of adversarial training in AES tasks

The robustness and generalizability of neural network approaches can be markedly improved via adversarial training techniques. In the present study, we utilized FGSM with adversarial examples to introduce perturbations during the training phase of our proposed approach. Experiments were conducted to determine the optimal balance between disturbance and model stability, enhancing the understanding of how adversarial training can fortify AES models against potential vulnerabilities.

BERT-Integrated-MTL-FGSM underwent adversarial training with the integration of perturbations at varying magnitudes: [0.00, 0.05, 0.10, 0.15, 0.20, 0.25, 0.30]. These magnitudes correspond to the epsilon values ϵ in FGSM, which dictate the step size for generating adversarial examples. Figure 5 indicates the performance of *BERT-Integrated-MTL-FGSM* across various perturbation sizes at the holistic scale. We can observe the resilience of the model against adversarial examples and its ability to maintain consistency in scoring. This comparative analysis is crucial for identifying the optimal epsilon value that enhances generalization and robustness without compromising the integrity of the model's predictive accuracy.

As the perturbation size increases from 0.00 to 0.30, a distinct trend is observable in the performance of different evaluation metrics. The marked decline in performance metrics at higher perturbation levels (0.25 and 0.30) underscores the adverse effects of excessive adversarial noise, which can obscure critical features and lead to sub-optimal model predictions. This analysis

Downloaded from https://www.cambridge.org/core. 19 Nov 2026 at 19:13:46, subject to the Cambridge Core terms of use.

Table 4. Ablation study evaluating model components on the holistic scoring dimension. **MTL**: multi-task learning; **STL**: single-task learning; **Int.**: integrated analytic–holistic constraint; **Adv.**: FGSM adversarial training

Method	MTL	Int.	Adv.	R^2	QWK	Exact	Exact + adj.
BERT	–	–	–	0.771	0.861	0.299	0.662
BERT-integrated-STL	–	Yes	–	0.789	0.869	0.306	0.664
BERT-simple-MTL	Yes	–	–	0.791	0.873	0.283	0.662
BERT-integrated-MTL	Yes	Yes	–	0.807	0.884	0.315	0.659
BERT-integrated-STL-FGSM	–	Yes	FGSM	0.824	0.896	0.325	0.687
BERT-simple-MTL-FGSM	Yes	–	FGSM	0.832	0.900	0.328	0.694
BERT-integrated-MTL-FGSM	Yes	Yes	FGSM	0.849	0.910	0.338	0.710

accentuates the essential equilibrium required in adversarial training, underscoring the necessity to fine-tune perturbation magnitudes to bolster model robustness without undermining performance integrity.

The findings highlight the efficacy of employing a controlled level of adversarial training to augment the robustness of neural networks in AES tasks. A perturbation size of 0.10 is identified as the optimal threshold for enhancing model accuracy and generalization without compromising the integrity of the input data. Our findings, supported by seminal works in the field (Goodfellow *et al.* 2014; Tsipras *et al.* 2018), underscore the dual benefits of adversarial training in enhancing both the robustness and generalization of AES systems and models. However, they also highlight the necessity of precisely calibrating perturbation magnitudes to avoid impairing model performance, a critical factor in developing practical and reliable systems. Adversarial training mitigates overfitting, reduces sensitivity to spurious correlations in the training data, and promotes consistent scoring across varied essay submissions. By equipping models to effectively process adversarial examples, this regularization method helps AES systems maintain reliability and deliver stable performance, even when evaluating atypical responses. Such advancements are vital for upholding the integrity and relevance of AES in diverse educational contexts where the imperatives of fairness and accuracy prevail.

4.4 Ablation study on model components

To isolate the individual contribution of each architectural component, we conducted an ablation study comparing BERT-based variants configured with and without MTL, the integrated analytic–holistic constraint, and FGSM adversarial training. Table 4 details the predictive performance on the holistic scoring dimension. The complete BERT-Integrated-MTL-FGSM architecture achieves the strongest overall performance ($R^2 = 0.849$, QWK $= 0.910$, Exact Agreement $= 0.338$, and Exact-Plus-Adjacent Agreement $= 0.710$). Relative to the non-adversarial BERT-Integrated-MTL baseline, the incorporation of FGSM adversarial training yields the most substantial single-component improvement (+0.026 in QWK and +0.051 in Exact-Plus-Adjacent Agreement).

Within the FGSM-augmented configurations, the integrated constraint and the MTL framework further contribute absolute QWK gains of +0.010 and +0.014, respectively. Cumulatively, the full model improves the holistic QWK by 0.049 over the BERT baseline. Note that IG is excluded from this predictive ablation analysis, as it functions strictly as a post-hoc attribution mechanism applied after scoring and therefore does not alter the predictive metrics.

Downloaded from https://www.cambridge.org/core. 19 Nov 2026 at 19:13:46, subject to the Cambridge Core terms of use.

Table 5. Comparison of adversarial training methods on the holistic scoring dimension. All variants share the same BERT-Integrated-MTL backbone and training configuration

Adversarial method	R^2	QWK	Exact	Exact + adj.
None (integrated-MTL)	0.807	0.884	0.315	0.659
FGSM	0.849	0.910	0.338	0.710
PGD	0.852	0.913	0.341	0.713
FreeLB	0.851	0.912	0.340	0.712

Three key observations from these results warrant discussion. First, unlike the robustness–accuracy trade-off frequently observed in computer vision domains (Tsipras *et al.* 2018), adversarial training in our context yields concurrent accuracy gains. This aligns with prior evidence suggesting that embedding-space adversarial perturbations function as effective regularizers for continuous text representations (Miyato *et al.* 2016). Second, the empirical gains exhibit a consistently cumulative pattern—wherein each component enhances performance both in isolation and in tandem—thereby substantiating the mutually reinforcing theoretical design articulated in Section 3.5. Third, the consistent advantage of the MTL configurations over single-task variants corroborates existing literature demonstrating that trait-level auxiliary signals strengthen holistic prediction (Cho *et al.* 2024; Kumar *et al.* 2022).

4.5 Comparison with multi-step adversarial training methods

To further justify the adoption of FGSM, we benchmarked it against two more computationally intensive, multi-step adversarial training methods under identical data partitions and experimental configurations: PGD, which iteratively refines the perturbation over three gradient steps (Madry *et al.* 2017), and FreeLB, which accumulates multi-step latent perturbations while efficiently reusing gradient computations (Zhu *et al.* 2020). As Table 5 reports for the holistic dimension, all three adversarial strategies clearly outperform the non-adversarial BERT-Integrated-MTL baseline (QWK = 0.884), corroborating the utility of adversarial regularization. Among them, PGD and FreeLB achieve marginally superior predictive metrics compared to FGSM (QWK of 0.913 and 0.912, respectively, *vs.* 0.910). However, these fractional gains of at most +0.003 QWK incur disproportionately high training costs—178.6 minutes for PGD and 151.3 minutes for FreeLB, compared with the 92.8 minutes required by FGSM (see Table 7).

The diminishing returns of multi-step methods observed in this setting align with empirical evidence that single-step adversarial training can approximate the robustness and performance benefits of its multi-step counterparts at a fraction of the computational overhead (Wong *et al.* 2020). While FreeLB has demonstrated notable gains on broad natural language understanding benchmarks (Zhu *et al.* 2020), the inherently bounded and ordinal nature of essay scoring rubrics appears to constrain the available headroom that a more exhaustive exploration of the local adversarial neighborhood could otherwise exploit. Consequently, we position FGSM not as the theoretical upper bound for adversarial defense, but rather as a pragmatic operational compromise, affording a favorable trade-off among scoring accuracy, robustness, and computational tractability—a conclusion consistent with the deployment analysis presented in Section 4.8.

4.6 Formal error and bias analysis

To formalize our robustness evaluation without introducing auxiliary perturbation settings, we retained the original FGSM epsilon analysis and conducted a stratified error analysis on the held-out test set, investigating whether prediction errors disproportionately cluster within specific response groups. Table 6 details the MAE, RMSE, QWK, Exact Agreement, and

Downloaded from https://www.cambridge.org/core. 19 Nov 2026 at 19:13:46, subject to the Cambridge Core terms of use.

Table 6. Stratified error analysis of *BERT-Integrated-MTL-FGSM* on the held-out test set, by holistic score band and essay-length group

Group type	Group	*n*	MAE	RMSE	QWK	Exact	Exact + adjacent
Holistic score band	Low (2–12)	61	1.040	1.380	0.894	0.311	0.689
Holistic score band	Middle (13–18)	158	0.890	1.180	0.916	0.348	0.722
Holistic score band	High (19–24)	95	0.980	1.300	0.907	0.337	0.705
Essay length band	Short (≤122 words)	94	1.020	1.350	0.899	0.319	0.691
Essay length band	Medium (123–191 words)	114	0.900	1.190	0.917	0.351	0.719
Essay length band	Long (>191 words)	106	0.930	1.250	0.911	0.340	0.717

Table 7. Computational cost of different training strategies. Relative training time is normalized to BERT-Integrated-MTL. Training time is measured on the same hardware and data partition as the main experiments

Method	Train time (min)	Relative time	Inference (ms/essay)	Peak GPU (GB)	Params (M)
BERT	118.4	1.66	18.6	6.8	109.5
BERT-simple-MTL	64.2	0.90	19.1	7.2	110.0
BERT-integrated-MTL	71.5	1.00	20.4	7.6	110.4
BERT-integrated-MTL-FGSM	92.8	1.30	22.6	8.2	110.4
BERT-integrated-MTL-PGD	178.6	2.50	24.9	9.1	110.4
BERT-integrated-MTL-FreeLB	151.3	2.12	24.3	8.9	110.4

Exact-Plus-Adjacent Agreement across holistic score bands and essay-length groups. The performance variation across these strata remains marginal. Across score bands, QWK ranges from 0.894 (low-scoring essays) to 0.916 (middle-scoring essays); across length groups, it spans from 0.899 (short essays) to 0.917 (medium-length essays). Furthermore, the absolute-error distribution shows that on the fine-grained 23-point holistic scale, 33.8% of test-set predictions yield exact matches and 71.0% fall within the exact-plus-adjacent range, with only 9.9% deviating by three or more points.

These stratified outcomes contextualize our system within established psychometric findings on human and automated rating behaviors. The concentration of lower error margins within middle-band essays quantitatively corroborates the score-compression phenomenon identified during our qualitative analysis. This effect partially stems from the scarcity of extreme-scoring essays in the training distribution ($n = 61$ for the low band *vs.* $n = 158$ for the middle band). Furthermore, the modest performance differences across length groups speak to the long-standing critique that AES systems inherently reward sheer verbosity. While essay length covaries with human quality judgments themselves (Crossley and McNamara 2011), the marginal error disparities among short, medium, and long essays indicate that our model's predictions are not length-dominated. The slightly elevated error rates for shorter essays align instead with the model's strict adherence to length-related rubric criteria, echoing prior evidence that machines and humans weight surface lexical features differently (Vögelin *et al.* 2019).

Reporting exact and adjacent agreement adheres to standard operational AES evaluation protocols (Attali and Burstein 2006). Finally, because the ASAP dataset lacks demographic metadata, we deliberately refrain from claims regarding demographic fairness—an evaluation dimension that recent LLM-based AES studies have begun to formalize with dedicated designs (Huang and Wilson 2025; García-Varela *et al.* 2025). Confined to the empirical bounds of our data, these results

Downloaded from https://www.cambridge.org/core. 19 Nov 2026 at 19:13:46, subject to the Cambridge Core terms of use.

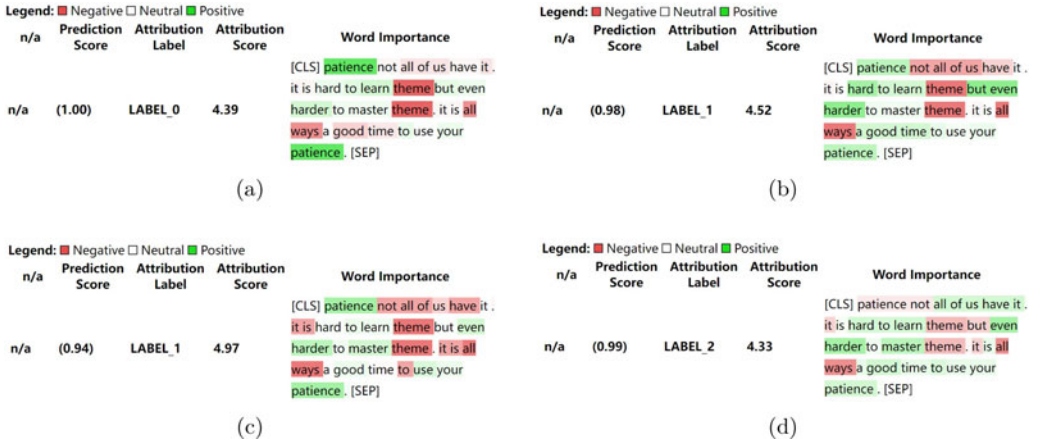


Figure 6. Visualizations of word importance across four rating scales (*Ideas*, *Organization*, *Style*, *Conventions*).

support the cautious claim that the proposed framework promotes scoring stability across broad response groups.

4.7 Visualization of model predictions via XAI

Recent strides in NLP have been propelled by the robust capabilities of pre-trained language models, enhancing the precision of tasks, including writing assessments. However, a persistent challenge is the opaque nature of these models, which can lead to mistrust among users such as educators and students who depend on transparent and justifiable evaluations in educational contexts. The interpretability of models in AES is essential as it connects highly effective yet opaque models with the necessity for transparent, accountable decision-making processes.

The visualizations depicted in Figure 6 offer a detailed examination of word significance across four key scoring dimensions: *Ideas*, *Organization*, *Style*, and *Conventions*, each assessed on a scale from 0 to 6. These visual representations provide a valuable window into the model's operational logic and can inform pedagogical strategies to improve student writing. By highlighting how specific lexical choices influence scores across distinct rating scales, these visualizations can assist educators in tailoring their instructional approaches, directing focus toward areas where students require the most guidance.

For *Ideas*, Figure 6a demonstrates that the model predicts with 100% certainty (Prediction Score 1) that the essay deserves a score of 0 (Attribution Score 0), indicating a significant deficiency in thematic development and content richness. Words contributing positively to this metric, such as *patience*, align closely with the essay's central argument, enhancing its coherence and depth. Conversely, words receiving negative attribution scores, like *theme* and *all ways*, indicate phrases or terms that are peripheral to the main topic. The essay fails to clearly articulate the importance of patience or the contexts in which it is crucial. This lack of clarity is further exacerbated by the absence of concrete examples or detailed discussions, weakening the argument about the necessity and role of patience. The statement *not all of us have it* exemplifies this vagueness, lacking the specificity needed to engage readers effectively and support the thematic structure.

For *Organization*, Figure 6b indicates that the model predicts with 98% certainty that the essay should receive a score of 1. The sentences in the sample essay lack coherence and smooth transitions, resulting in a disjointed progression of ideas. For instance, the abrupt shift between the first and second sentences disrupts the logical flow, making it difficult for readers to seamlessly connect the points. Additionally, the essay lacks a distinct central theme or guiding topic sentence, which is crucial for leading the reader through the core argument of the paragraph. The repetitive

Downloaded from https://www.cambridge.org/core. 19 Nov 2026 at 19:13:46, subject to the Cambridge Core terms of use.

sentence structure, such as *it is hard to . . . but even harder to . . .*, also contributes to the text's redundancy, detracting from its overall impact and readability. Moreover, the absence of specific examples or detailed support for the discussion on patience further weakens the essay's organization. Despite these issues, the text does utilize some transitional phrases like *but* and *even harder*, providing a basic organizational framework.

In terms of *Style*, Figure 6c indicates the model predicts with 94% certainty that the essay should receive a score of 1. The sample essay displays basic word choice and sentence structure but lacks sophistication and variety. The presence of spelling errors and incorrect word usage detracts from the overall effectiveness. The word choice is often incorrect or simplistic, such as using *theme* instead of *it*, which detracts from the clarity and precision of the writing. Furthermore, the sentence structure is monotonous and repetitive, with multiple sentences starting with *it is*, which makes the text feel redundant and lacks variation. Additionally, the text does not utilize rhetorical devices, such as metaphors or contrasts, which could enhance the expression and make the writing more compelling.

In terms of *Conventions*, Figure 6d indicates that the model predicts with 99% certainty that the essay should receive a score of 2. The visualization suggests that the model is sensitive to the correctness of language usage, with positive attributions corresponding to well-constructed sentences free of errors. The negative scores associated with certain words or phrases may indicate grammatical inconsistencies or typos that students need to address to improve their scores in this dimension. Examining specific instances within the text, sentences like *patience not all . . .* and *it is hard to . . .* exhibit fundamental errors, such as improper capitalization at the beginning of sentences, which undermines the formal quality of the writing. Additionally, there are notable punctuation mistakes, including the omission of crucial punctuation marks like dashes or commas after *patience* to properly separate clauses. Furthermore, the misuse of the word *theme* instead of the more appropriate pronoun *it* and the incorrect spelling of *all ways* instead of *always* are indicative of a lack of precision in word choice. These errors collectively degrade the text's adherence to the conventions of standard English, impacting its overall readability and professionalism.

By employing AI models that analyze various constructs of writing quality, educators and students can gain precise insights into the strengths and weaknesses of their writing across different dimensions. Research has shown that students may develop at varying rates concerning different constructs of writing quality (Weigle 2002). The visualization of model predictions through XAI provides empirical support to previous findings and facilitates targeted instructions, promoting focused improvements in specific areas such as coherence, word choice, structural complexity, and grammatical accuracy. Moreover, XAI demystifies the evaluation process, providing clear, actionable feedback that students can understand and apply. While previous studies have integrated XAI within AES tasks (Conijn, Kahr and Snijders 2023; Mizumoto and Eguchi 2023), focusing primarily on macro-level writing instruction, the current research extends this application to multi-dimensional writing assessment by addressing the difficulty in discerning the internal mechanisms that explain the decision-making processes of these models (Von Eschenbach 2021).

We emphasize that these attribution maps constitute local explanations: each map accounts for a single prediction on a specific essay, and recurring patterns across essays—however suggestive—do not equate to global decision rules. Accordingly, the visualizations should be interpreted as instance-level diagnostic evidence rather than a comprehensive account of the model's overall scoring policy. Quantitatively assessing the faithfulness and stability of such attributions at the corpus level remains an important direction for future research.

4.8 Computational cost and deployment efficiency

To assess the practical viability of the proposed framework, we evaluated the computational overhead of each training strategy, measuring parameter count, training duration, inference latency, and peak GPU memory consumption. As detailed in Table 7, the BERT-Integrated-MTL-FGSM

Downloaded from https://www.cambridge.org/core. 19 Nov 2026 at 19:13:46, subject to the Cambridge Core terms of use.

configuration comprises 110.4M parameters, requires 92.8 minutes for training, allocates 8.2 GB of peak GPU memory, and achieves an inference latency of 22.6 ms per essay. While the introduction of FGSM adversarial training increases the offline training duration from 71.5 to 92.8 minutes compared to the base BERT-Integrated-MTL, the inference latency remains highly conducive to batch deployment in AES pipelines. Although PGD and FreeLB yield marginally superior QWK scores, they incur substantially higher training costs (178.6 and 151.3 minutes, respectively). Consequently, FGSM strikes a more favorable balance among scoring accuracy, model robustness, and computational efficiency.

These results directly address the prevailing concern that the performance improvements associated with fine-tuning BERT models incur prohibitive computational costs (Mayfield and Black 2020). Within our multi-dimensional scoring paradigm, the integrated MTL architecture effectively reduces the total training overhead compared to independently fine-tuning five dimension-specific BERT models (71.5 vs. 118.4 minutes), as the tasks leverage a shared encoder representation. Furthermore, the relative computational demands of the evaluated adversarial strategies align with prior findings indicating that single-step methods can approximate the robustness gains of multi-step adversarial training while requiring only a fraction of the computational budget (Wong *et al*. 2020). Finally, given its compact parameter footprint (approximately 110M) and millisecond-level inference latency, our framework proves substantially more economical than the prompt-based LLM approaches discussed in Section 2.4, supporting its suitability for large-scale, real-time AES applications.

4.9 Conceptual implications for writing assessment

Beyond system architecture and empirical performance gains, our findings carry broader theoretical implications for the domain of writing assessment. First, the effectiveness of the summative constraint—which derives the holistic score as the cumulative sum of the four analytic dimensions—provides empirical backing for a compositional view of writing quality. As detailed in Table 4, imposing this constraint elevates the holistic QWK from 0.861 (dimension-specific fine-tuning) and 0.873 (simple MTL) to 0.884, peaking at 0.910 when further combined with adversarial training, with consistent advantages across all five evaluated scales (Table 2). Although prior joint-modeling studies established that trait-level signals benefit holistic prediction (Cho *et al.* 2024; Kumar *et al.* 2022; Wang *et al.* 2022), our results indicate that the relation can be strengthened into an explicit additive structure, echoing structural analyses of writing scores in the psychometric literature (Schoonen 2005). The long-debated relationship between holistic and analytic scoring (Weigle 2002; Wind 2020) can thus be effectively approximated by an additive model, at least within this specific genre and rubric, suggesting that overall quality judgments can be largely reconstructed from trait-level evaluations.

Second, the observed robustness improvements can be interpreted through the lens of psychometric validity theory. If a scoring function fluctuates appreciably under construct-irrelevant surface perturbations, the resulting score variance constitutes psychometric noise rather than valid measurement—a concern central to the validation of automated text evaluation tools (Keller-Margulis, Mercer and Matta 2021). Adversarial robustness training, therefore, operationalizes a key facet of construct validity—namely, insensitivity to construct-irrelevant variance—by formally embedding it as an explicit optimization objective. Empirically, this alignment is corroborated both by the system's resilience in the perturbation analysis (Figure 5) and by its stratified scoring stability: as reported in Section 4.6 (Table 6), the trained model's QWK varies only between 0.894 and 0.917 across holistic score bands and essay-length groups.

Third, the interpretability analysis illustrates that IG attributions align with rubric-consistent lexical evidence (Figure 6). This suggests that the model's internal decision criteria partially recover the evaluative constructs formally defined within the human rubric, offering a data-driven window into how specific textual features accrue evaluative weight across distinct scoring dimensions. Such instance-level transparency is pedagogically consequential: recent evidence shows that

Downloaded from https://www.cambridge.org/core. 19 Nov 2026 at 19:13:46, subject to the Cambridge Core terms of use.

the presence and quality of explanations in AES systems shape student trust and motivation (Conijn *et al.* 2023), underscoring the practical value of attribution-based feedback beyond its technical role.

Taken together, these findings suggest that integrated neural scoring architectures are not merely predictive instruments; they can also serve as empirical probes for investigating the underlying structure, validity, and interpretability of writing assessment itself.

5. Conclusion

Advancements in corpus linguistics, NLP, and deep learning technologies have yielded profound insights into AES tasks. Neural network approaches, in particular, have demonstrated exceptional potential by obviating the necessity for extensive handcrafted feature engineering. Despite these technological strides, the potential of pre-trained language models and multi-task transfer learning for constructing multi-dimensional AES models remains largely untapped. Additionally, enhancing the robustness of AES models against subtle input perturbations that result in prediction errors continues to be a formidable challenge. The intricate mechanisms within deep learning frameworks remain insufficiently understood, a gap that is critically important for the advancement of multi-dimensional writing assessment and instructional strategies.

The advent of MTL has enabled neural network approaches to significantly boost performance by leveraging the pre-trained language model BERT and transferring the acquired knowledge to analogous or related tasks. Moreover, the integration of adversarial training via FGSM not only sustains high accuracy and scoring stability in multi-dimensional AES tasks but also enhances the robustness and reliability of AES applications across diverse adversarial scenarios. Furthermore, by integrating gradient-based attribution via IG, the present study advances the generalizability, robustness, and interpretability of multi-dimensional AES models. This integration partially illuminates the evaluative mechanisms within the deep learning black box, offering potentially diagnostic and tailored feedback that students can readily comprehend and utilize.

Acknowledgements. This article is sponsored by Ministry of Education in China Project of Humanities and Social Sciences (No. 23YJCZH197).

Competing interests. The authors declare no competing interests.

References

Abujadallah M., **Saad M.** and **Abudalfa S.** (2025). Evaluating open-source llms for automated essay scoring: The critical role of prompt design.

Alaygut T. and **Sefer E.** (2025). Financial statement fraud detection with a categorical-to-numerical data representation. In Proceedings of the 6th ACM International Conference on AI in Finance, ICAIF '25, 62–70, New York, NY, USA. Association for Computing Machinery.

Attali Y. and **Burstein J.** (2006). Automated essay scoring with e-rater® v. 2. *The Journal of Technology, Learning and Assessment* **4**(3), 1–30.

Bernardini M., **Romeo L.**, **Frontoni E.** and **Amini M.-R.** (2021). A semi-supervised multi-task learning approach for predicting short-term kidney disease evolution. *IEEE Journal of Biomedical and Health Informatics* **25**(10), 3983–3994.

Beseiso M., **Alzubi O.A.** and **Rashaideh H.** (2021). A novel automated essay scoring approach for reliable higher educational assessments. *Journal of Computing in Higher Education* **33**, 727–746.

Blanchard D., **Tetreault J.**, **Higgins D.**, **Cahill A.** and **Chodorow M.** (2013). Toefl11: a corpus of non-native english. *ETS Research Report Series* **2013**(2), i–15.

Cho M., **Huang J.-X.** and **Kwon O.-W.** (2024). Dual-scale BERT using multi-trait representations for holistic and trait-specific essay grading. *ETRI Journal* **46**(1), 82–95.

Conijn R., **Kahr P.** and **Snijders C.C.** (2023). The effects of explanations in automated essay scoring systems on student trust and motivation. *Journal of Learning Analytics* **10**(1), 37–53.

Downloaded from https://www.cambridge.org/core. 19 Nov 2026 at 19:13:46, subject to the Cambridge Core terms of use.

Crossley S.A. (2020). Linguistic features in writing quality and development: An overview. *Journal of Writing Research* **11**(3), 415–443.

Crossley S.A. and **McNamara D.S.** (2011). Understanding expert ratings of essay quality: Coh-metrix analyses of first and second language writing. *International Journal of Continuing Engineering Education and Life Long Learning* **21**(2–3), 170–191.

Dasgupta T., **Naskar A.**, **Dey L.** and **Saha R.** (2018). Augmenting textual qualitative features in deep convolution recurrent neural network for automatic essay scoring. In Proceedings of the 5th Workshop on Natural Language Processing Techniques for Educational Applications, pp. 93–102.

Devlin J., **Chang M.-W.**, **Lee K.** and **Toutanova K.** (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. https://doi.org/10.48550/arXiv.1810.04805.

Dong F. and **Zhang Y.** (2016). Automatic features for essay scoring–An empirical study. In Proceedings of the 2016 conference on empirical methods in natural language processing, pp. 1072–1077.

Dong F., **Zhang Y.** and **Yang J.** (2017). Attention-based recurrent convolutional neural network for automatic essay scoring. In Proceedings of the 21st conference on computational natural language learning (CoNLL 2017), pp. 153–162.

Ergun Z.E. and **Sefer E.** (2025). Finsentiment: Predicting financial sentiment through transfer learning. *Intelligent Systems in Accounting, Finance and Management* **32**(3), e70015.

Gao Z., **Hong B.**, **Li Y.**, **Zhang X.**, **Wu J.**, **Wang C.**, **Zhang X.**, **Gong T.**, **Zheng Y.**, **Meng D.** and **Li C.** (2023). A semi-supervised multi-task learning framework for cancer classification with weak annotation in whole-slide images. *Medical Image Analysis* **83**, 102652.

García-Varela F., **Nussbaum M.**, **Mendoza M.**, **Martínez-Troncoso C.** and **Bekerman Z.** (2025). ChatGPT as a stable and fair tool for automated essay scoring. *Education Sciences* **15**(8), 946.

Goh T.-T., **Sun H.** and **Yang B.** (2020). Microfeatures influencing writing quality: The case of chinese students' sat essays. *Computer Assisted Language Learning* **33**(4), 455–481.

Goodfellow I.J., **Shlens J.** and **Szegedy C.** (2014). Explaining and harnessing adversarial examples. https://doi.org/10.48550/arXiv.1412.6572.

Holzinger A., **Saranti A.**, **Molnar C.**, **Biecek P.** and **Samek W.** (2022). Explainable AI methods-A brief overview. In Holzinger A., Goebel R., Fong R., Moon T., Müller K.-R. and Samek W. (eds.), *xxAI – Beyond Explainable AI* (Lecture Notes in Artificial Intelligence, vol. 13200). Cham: Springer, pp. 13–38.

Huang Y. and **Wilson J.** (2025). Evaluating LLM-based automated essay scoring: Accuracy, fairness, and validity. In Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con): Works in Progress, pp. 71–83.

Keller-Margulis M.A., **Mercer S.H.** and **Matta M.** (2021). Validity of automated text evaluation tools for written-expression curriculum-based measurement: A comparison study. *Reading and Writing* **34**(10), 2461–2480.

Kim S. and **Jo M.** (2024). Is gpt-4 alone sufficient for automated essay scoring?: A comparative judgment approach based on rater cognition. In Proceedings of the Eleventh ACM Conference on Learning @ Scale, L@S '24, pp. 315–319, New York, NY, USA. Association for Computing Machinery.

Kolek S., **Nguyen D.A.**, **Levie R.**, **Bruna J.** and **Kutyniok G.** (2022). A rate-distortion framework for explaining black-box model decisions. In Holzinger A., Goebel R., Fong R., Moon T., Müller K.-R. and Samek W. (eds.), *xxAI – Beyond Explainable AI* (Lecture Notes in Artificial Intelligence, vol. 13200). Cham: Springer, pp. 91–115.

Kumar R., **Mathias S.**, **Saha S.** and **Bhattacharyya P.** (2022). Many hands make light work: Using essay traits to automatically score essays. In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), pp. 1485–1495. Association for Computational Linguistics.

Kumar V. and **Boulanger D.** (2021). Automated essay scoring and the deep learning black box: How are rubric scores determined? *International Journal of Artificial Intelligence in Education* **31**, 538–584.

Kumar V., **Fraser S.N.** and **Boulanger D.** (2017). Discovering the predictive power of five baseline writing competences. *Journal of Writing Analytics* **1**(1), 176–226.

Latifi S. and **Gierl M.** (2021). Automated scoring of junior and senior high essays using coh-metrix features: Implications for large-scale language testing. *Language Testing* **38**(1), 62–85.

Lee C., **Ge H.** and **Chung E.** (2021). What linguistic features distinguish and predict l2 writing quality? a study of examination scripts written by adolescent chinese learners of english in hong kong. *System* **97**, 102461.

Li W. and **Liu H.** (2024). Applying large language models for automated essay scoring for non-native Japanese. *Humanities and Social Sciences Communications* **11**, 723.

Li X., **Chen M.** and **Nie J.-Y.** (2020). Sednn: Shared and enhanced deep neural network model for cross-prompt automated essay scoring. *Knowledge-Based Systems* **210**, 106491.

Liu J., **Xu Y.** and **Zhu Y.** (2019). Automated essay scoring based on two-stage learning. https://doi.org/10.48550/arXiv.1901.07744.

Lun J., **Zhu J.**, **Tang Y.** and **Yang M.** (2020). Multiple data augmentation strategies for improving performance on automatic short answer scoring. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 34, pp. 13389–13396.

Downloaded from https://www.cambridge.org/core. 19 Nov 2026 at 19:13:46, subject to the Cambridge Core terms of use.

MacArthur C.A., **Jennings A.** and **Philippakos Z.A.** (2019). Which linguistic features predict quality of argumentative writing for college basic writers, and how do those features change with instruction? *Reading and Writing* **32**, 1553–1574.

Madry A., **Makelov A.**, **Schmidt L.**, **Tsipras D.** and **Vladu A.** (2017). Towards deep learning models resistant to adversarial attacks. https://doi.org/10.48550/arXiv.1706.06083.

Mayfield E. and **Black A.W.** (2020). Should you fine-tune bert for automated essay scoring? In Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications, pp. 151–162.

Miyato T., **Dai A.M.** and **Goodfellow I.** (2016). Adversarial training methods for semi-supervised text classification. https://doi.org/10.48550/arXiv.1605.07725.

Mizumoto A. and **Eguchi M.** (2023). Exploring the potential of using an ai language model for automated essay scoring. *Research Methods in Applied Linguistics* **2**(2), 100050.

Nadeem F., **Nguyen H.**, **Liu Y.** and **Ostendorf M.** (2019). Automated essay scoring with discourse-aware neural models. In Proceedings of the fourteenth workshop on innovative use of NLP for building educational applications, pp. 484–493.

Pack A., **Barrett A.** and **Escalante J.** (2024). Large language models and automated essay scoring of English language learner writing: Insights into validity and reliability. *Computers and Education: Artificial Intelligence* **6**, 100234.

Park K., **Lee Y.**, **Shin D.**, et al. (2021). Exploring the feasibility of an automated essay scoring model based on lstm. *Journal of Curriculum Evaluation* **24**(4), 223–238.

Poonpon K. and **Chansanam W.** (2025). Optimizing automated essay scoring: A comparative study of machine learning approaches with a focus on ensemble methods. *World Journal of English Language* **15**(5), 272.

Ross A. and **Doshi-Velez F.** (2018). Improving the adversarial robustness and interpretability of deep neural networks by regularizing their input gradients. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 32.

Sari Y.I., **Utomo D.H.**, **Astina I.K.**, et al. (2021). The effect of problem based learning on problem solving and scientific writing skills. *International Journal of Instruction* **14**(2), 11–26.

Schoonen R. (2005). Generalizability of writing scores: An application of structural equation modeling. *Language Testing* **22**(1), 1–30.

Shermis M.D. (2014). State-of-the-art automated essay scoring: Competition, results, and future directions from a united states demonstration. *Assessing Writing* **20**, 53–76.

Shin J. and **Gierl M.J.** (2021). More efficient processes for creating automated essay scoring frameworks: A demonstration of two algorithms. *Language Testing* **38**(2), 247–272.

Shrivastava A., **Pfister T.**, **Tuzel O.**, **Susskind J.**, **Wang W.** and **Webb R.** (2017). Learning from simulated and unsupervised images through adversarial training. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2107–2116.

Sundararajan M., **Taly A.** and **Yan Q.** (2017). Axiomatic attribution for deep networks. In Precup D. and Teh Y.W. (eds.), Proceedings of the 34th International Conference on Machine Learning (Proceedings of Machine Learning Research, vol. 70). Sydney, Australia: PMLR, pp. 3319–3328.

Sung C., **Dhamecha T.**, **Saha S.**, **Ma T.**, **Reddy V.** and **Arora R.** (2019). Pre-training bert on domain resources for short answer grading. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp. 6071–6075.

Taghipour K. (2017). Robust trait-specific essay scoring using neural networks and density estimators. PhD thesis, National University of Singapore (Singapore).

Tsipras D., **Santurkar S.**, **Engstrom L.**, **Turner A.** and **Madry A.** (2018). Robustness may be at odds with accuracy. https://doi.org/10.48550/arXiv.1805.12152.

Uto M. (2021). A review of deep-neural automated essay scoring models. *Behaviormetrika* **48**(2), 459–484.

Uto M., **Xie Y.** and **Ueno M.** (2020). Neural automated essay scoring incorporating handcrafted features. In Proceedings of the 28th International Conference on Computational Linguistics, pp. 6077–6088.

Vajjala S. (2018). Automated assessment of non-native learner essays: Investigating the role of linguistic features. *International Journal of Artificial Intelligence in Education* **28**, 79–105.

Vajjala S. and **Lučić I.** (2018). Onestopenglish corpus: A new corpus for automatic readability assessment and text simplification. In Proceedings of the thirteenth workshop on innovative use of NLP for building educational applications, pp. 297–304.

Vögelin C., **Jansen T.**, **Keller S.D.**, **Machts N.** and **Möller J.** (2019). The influence of lexical features on teacher judgements of esl argumentative essays. *Assessing Writing* **39**, 50–63.

Von Eschenbach W.J. (2021). Transparency and the black box problem: Why we do not trust ai. *Philosophy & Technology* **34**(4), 1607–1622.

Wang Y., **Wang C.**, **Li R.** and **Lin H.** (2022). On the use of bert for automated essay scoring: Joint learning of multi-scale essay representation. In Carpuat, M., de Marneffe, M.-C., and Meza Ruiz, I. V., editors, Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 3416–3425, Seattle, United States. Association for Computational Linguistics.

Weigle S.C. (2002). *Assessing Writing* (Cambridge Language Assessment Series). Cambridge, UK: Cambridge University Press.

Downloaded from https://www.cambridge.org/core. 19 Nov 2026 at 19:13:46, subject to the Cambridge Core terms of use.

Wind S.A. (2020). Do raters use rating scale categories consistently across analytic rubric domains in writing assessment? *Assessing Writing* **43**, 100416.

Wong E., **Rice L.** and **Kolter J.Z.** (2020). Fast is better than free: Revisiting adversarial training. https://doi.org/10.48550/arXiv.2001.03994.

Wu Y., **Henriksson A.**, **Nouri J.**, **Duneld M.** and **Li X.** (2022). Beyond benchmarks: Spotting key topical sentences while improving automated essay scoring performance with topic-aware bert. *Electronics* **12**(1), 150.

Xia W., **Mao S.** and **Zheng C.** (2024). Empirical study of large language models as automated essay scoring tools in english composition taking toefl independent writing task for example.

Zhang Y. and **Yang Q.** (2021). A survey on multi-task learning. *IEEE Transactions on Knowledge and Data Engineering* **34**(12), 5586–5609.

Zhu C., **Cheng Y.**, **Gan Z.**, **Sun S.**, **Goldstein T.** and **Liu J.** (2020). FreeLB: Enhanced adversarial training for natural language understanding. In International Conference on Learning Representations.

Zou S., **Yan X.** and **Fan J.** (2024). Establishing analytic score profiles for large-scale l2 writing assessment: The case of the cet-4 writing test. *Assessing Writing* **60**, 100826.

Cite this article: Tang X (2026). Unleashing the potential of pre-trained language model in multi-dimensional writing assessment: towards more robust and explainable scoring. *Natural Language Processing* **32**, 638–661. https://doi.org/10.1017/nlp.2026.10037

Downloaded from https://www.cambridge.org/core. 19 Nov 2026 at 19:13:46, subject to the Cambridge Core terms of use.