SCIENTIFIC CULTURE
www.sci-cult.com

SCIENTIFIC CULTURE, Vol. 12, No. 4, (2026), pp. 1474-1489
Open Access. Online & Print

DOI: 10.5281/zenodo.12426230

ADVERSARIAL DEBIASING FOR FAIR AUTOMATED ESSAY SCORING ACROSS LINGUISTIC BACKGROUNDS

Rowaidah Al Abdullah [1*], Fatma Al Shamli [2], Amal Al Abri [3], Sara Al Moqbali [4], and Nasim Al Balushi [5]

[1] E-Learning Services Section & General Foundation Department, Management Information System, Mazoon College, Muscat, Oman.
[2] Computing and Information Sciences Department, College of Computing and Information Sciences, Database, University of Technology and Applied Science, Muscat, Oman.
[3] Computing and Information Sciences Department, College of Computing and Information Sciences, Deep learning, University of Technology and Applied Science, Muscat, Oman.
[4] Computing and Information Sciences Department, College of Computing and Information Sciences, Data science, University of Technology and Applied Science, Muscat, Oman.
[5] Computing and Information Sciences Department, College of Computing and Information Sciences, Networking, University of Technology and Applied Science, Muscat, Oman.

Received: 17/07/2025
Accepted: 29/01/2026

Corresponding author: Rowaidah Al Abdullah (ruwaida.alabdullah@mazcol.edu.om)

ABSTRACT

Automated Essay Scoring (AES) systems are widely used in educational assessment; however, there are still certain issues related to demographic biases, in particular, gender- and language-related biases. In this paper, a new model called FairGrade is proposed to reduce these biases without compromising scoring accuracy. FairGrade is based on the architecture of BERT and adds an adversarial discriminator that has a Gradient Reversal Layer (GRL), which aims to reduce the influence of demographic features on scoring predictions. The model optimizes the BERT encoder to obtain salient features in essays, and a multi-task learning structure optimizes both essay scoring and mitigation of unfairness. We tested FairGrade on two datasets, including the ASAP dataset, which contained native English essays, and the TOEFL11 dataset, which contained non-native English essays. These findings indicate that FairGrade achieves higher accuracy and improved equity than baseline models. It improves Quadratic Weighted Kappa (QWK) by 3.7% over the BERT baseline on the ASAP dataset and increases classification accuracy by 4.3% on the TOEFL11 dataset. Furthermore, the model reduces the performance gap between language groups by 50.4%. The adversarial debiasing mechanism substantially reduces gender and L1-related biases, reducing gender bias by 78.1% and L1-related bias by 50.4%. Such results suggest that fairness-aware AES models, including FairGrade, can both improve prediction quality and equity in educational tests and thus promote a more inclusive assessment paradigm.

KEYWORDS: Automated Essay Scoring (AES), Fairness, Adversarial Debiasing, BERT, Linguistic Bias, Educational Assessment.

1. INTRODUCTION

Automated Essay Scoring (AES) systems are already a part of large-scale educational testing, including high-stakes tests like the Test of English as a Foreign Language (TOEFL), the Graduate Record Examination (GRE), and other university admissions tests (Zhong et al., 2024). These machine-learning-based systems offer significant benefits, including efficiency, consistency, and scalability.

Large-scale educational assessment has increasingly adopted Automated Essay Scoring (AES) systems (Williamson, Xi, & Breyer, 2012). AES performance has improved significantly with the emergence of neural architectures (Taghipour & Ng, 2016; Dong et al., 2017). Recent studies indicate a shift toward transformer-based AES models (Ke & Ng, 2019; Shi and Aryadoust, 2022).

Nevertheless, despite these technological improvements, the growing adoption of AES systems has led to debates about fairness, especially in linguistically and culturally diverse settings, where demographic biases can be reinforced (Schaller et al., 2024). Like most AI-based models, AES systems are prone to biases in the training data (Litman et al., 2021). Prior research indicates that AES systems may encode and reproduce demographic differences, which results in unfair judgments. As an example, gender biases have been identified, whereby the essays submitted by female students receive lower grades as compared to those submitted by male students of the same quality (Andersen et al., 2025; Latif, Zhai, and Liu, 2023). Likewise, linguistic biases can also be observed, with essays of non-native English speakers usually scoring lower than the ones of native writers with similar quality of content (Fan and Yun, 2025). These differences compromise the quality of standardized tests and can be disproportionately applied to underrepresented student groups.

Although the question of bias in AES systems has been studied, much of the available literature has been more concerned with enhancing the accuracy of scoring, without paying much attention to the issue of fairness (Zhong et al., 2024). Even though there are studies that admit that there are demographic differences, they usually use post-hoc analyses instead of considering fairness constraints in the process of training the model (Schaller et al., 2024; Litman et al., 2021). This disjuncture indicates the necessity of methodological solutions that would incorporate fairness into the learning process, not as a secondary modification (Ferrara et al., 2024).

To overcome this issue, this paper presents FairGrade, a new deep learning architecture that can reduce both gender and linguistic bias in AES systems without compromising predictive accuracy. By introducing adversarial debiasing as a part of the training pipeline. FairGrade reduces demographic bias and preserves robust scoring accuracy.

The rest of this paper is structured in the following way. Section 2 presents the literature review of the previous studies on the Automated Essay Scoring (AES), with a special focus on the demographic bias and the fairness-conscious models. Section 3 outlines the suggested methodology, comprising of the datasets, preprocessing, FairGrade architecture, adversarial debiasing mechanism, training strategy, and evaluation measures. Section 4 reports the experiment findings, such as performance comparison, fairness analysis, ablation test, and cross-dataset generalized experiments. Section 5 explains the implication of the findings and the strength factors and disadvantages of the current approach. Lastly, Section 6 is a conclusion to the paper and provides future research directions.

This paper makes the following key contributions:

1. Adversarial debiasing framework for AES: We propose a Gradient Reversal Layer (GRL)-based adversarial debiasing approach to mitigate demographic bias during the training of Automated Essay Scoring (AES) systems, while maintaining predictive performance.
2. Dual-dataset evaluation: FairGrade is evaluated on two complementary datasets: the ASAP dataset (12,976 native English essays) and the TOEFL11 dataset (12,100 essays across 11 first-language backgrounds), enabling assessment across diverse linguistic contexts.
3. Comprehensive fairness analysis: The study evaluates fairness using multiple metrics, including Demographic Parity, Equalized Odds, and classifier-based auditing, providing a detailed assessment of bias mitigation.
4. Cross-dataset generalizability: Results indicate that the debiased representations learned by FairGrade generalize across datasets, maintaining both fairness and scoring performance across varied linguistic settings.

2. RELATED WORKS

Automatic essay scoring (AES) systems have undergone significant development over the last several decades, driven by advances in natural language processing (NLP) and machine learning to provide scalable educational evaluation. Previously,

AES technologies were based on the intensive use of hand-crafted linguistic and surface-level cues, including the number of grammatical objects, vocabulary richness measures, and indicators of syntactic complexity (Cummins et al., 2016). Discourse-modeling approaches further improved coherence evaluation by reflecting the higher-order organizational structures (Persing & Ng, 2015).

With advanced neural designs, the development of deep learning has significantly enhanced the performance in essay-scoring with a substantial improvement in scores (Taghipour & Ng, 2016; Dong et al., 2017). Methods based on transformers have also enhanced the strength of AES and contextual modeling (Mayfield & Black, 2020). Particularly, long-document transformers allow to effectively process long essays, thus alleviating the memory pressure of traditional transformer architectures (Beltagy et al., 2020). Recent surveys highlight the greater movement towards transformer-based AES models and contextual representation learning (Ke & Ng, 2019; Shi & Aryadoust, 2022). Jong et al. (2023), in their turn, address the importance of feedback as one of the key elements in the context of AES and state that these platforms often lack such a feature, and such a lack can worsen the existing disparities.

Despite these technological advancements, the issue of fairness has become even more relevant. The bias of AI in education by algorithms has been widely reported (Baker & Hawn, 2022). Automated scoring systems have also presented issues of fairness, especially in subgroups of demographic diversity (Litman et al., 2021; Schaller et al., 2024). Even though the use of reinforcement-learning methods was studied to improve scoring accuracy in AES, those methods, in most cases, fail to explicitly resolve fairness limitations (Zhang et al., 2018).

Gender and language biases remain persistent challenges in AES and NLP in general. As a systematic difference, NLP systems can be biased at the level of linguistic communities, particularly in low-resource or underrepresented communities (Blodgett et al., 2020; Joshi et al., 2020). Gender stereotypes are encoded in word-embedding models (Bolukbasi et al, 2016), and gender bias has been observed to persist in downstream NLP tasks like coreference resolution and language modeling (Zhao et al, 2018). In addition, post-hoc debiasing techniques often do not remove all demographic cues in the learned representations (Gonen & Goldberg, 2019).

In addition to feature engineering and post-processing corrections, an effective in-processing method, adversarial debiasing, has been demonstrated to be useful in the reduction of bias during the training of models. Zhang et al. (2018) proposed adversarial learning paradigms to reduce demographic signal leakage during representation learning. This method has been recently improved, and the studies have shown its effectiveness in reducing implicit biases that are added in preprocessing and model optimization (Cheng et al., 2024).

Collectively, these contributions indicate that there are major gaps in existing studies. Much of the current AES literature focuses on scoring accuracy and feedback quality, while fairness considerations are relegated to the background or evaluated only post hoc. Although fairness audits and post hoc corrections mitigate bias, they do not directly address biases embedded in learned representations. FairGrade can resolve this drawback by introducing adversarial debiasing as part of the AES training pipeline, thereby reducing both gender and linguistic biases without affecting predictive performance. This direction contributes to the creation of equity-aware AES systems in line with the growing need for just educational technologies.

3. METHODOLOGY

3.1. Datasets

The ASAP dataset includes 12,976 essays written by native English-speaking students in 7-10 grades. This corpus is heterogeneous regarding the essay prompts, including persuasive, source-based, and narrative ones. The challenges presented by each prompt make the dataset especially appropriate in assessing the Automated Essay Scoring (AES) systems in various writing genres of academic writing.

The Quadratic Weighted Kappa (QWK) is used as the main evaluation tool to evaluate the performance of the AES system. QWK, unlike traditional measures of accuracy, takes into consideration the ordinal character of essay scores, which gives a more accurate measure of model performance. This is especially relevant to AES systems, in which a slight difference in the quality of an essay can significantly affect the scoring results.

The TOEFL11 dataset, in turn, includes 12,100 non-native English essays, which were collected among eleven different first-language (L1) groups, such as Arabic, Chinese, Hindi, and Telugu, among others. This dataset aims to evaluate the performance of the AES system faced with essays by

students with different degrees of English proficiency. Essays are labeled as Low, Medium, or High.

Linguistic bias is a major problem in the TOEFL11 dataset, where AES systems are likely to show worse results on non-native English essays when trained on native English data. This difference is mainly caused by the variation in writing style, grammar, and vocabulary between non-native and native speakers. The TOEFL11 dataset can be used to investigate accuracy disparities among different L1 groups and emphasize the ways in which language backgrounds can be disadvantaged in the process of scoring.

Both the ASAP and TOEFL11 corpora are subjected to necessary preprocessing before they are fed into the FairGrade model. The essays are first tokenized using the Word Piece tokenizer that breaks the text into sub word units. The approach improves effective management of infrequent words and morphological variations. Each essay is limited to 512 tokens, which is the maximum sequence length supported by BERT, which is in line with the needs of BERT-based models. Essays with fewer than 512 tokens are padded to make them have the same size of input across all data points. Lastly, normalization processes are used, which include changing all text to lower case and normalizing whitespace while preserving punctuation. These preprocessing procedures normalize the datasets, thus allowing the model to manipulate the inputs well to reduce the unnecessary data that could otherwise hinder the learning process.

3.2. FairGrade Architecture

The FairGrade model is built on BERT architecture with certain changes aimed at improving the performance of the essay grading system, as well as reducing demographic biases at the same time. The architecture consists of two major components, namely the Grade Predictor and Adversarial Discriminator, as shown in Figure 1.

FairGrade uses the BERT-base-uncased encoder, which is a twelve-layer pre-trained model with 768 hidden dimensions, first trained on a large, general-domain corpus. This encoder is trained on an essay corpus, and its goal is to enhance the precision of scoring essays as well as the equity of the model's prediction. In the fine-tuning step, the last four layers of the transformer are retrained, and previous layers are fixed, thus retaining the general linguistic information learned during pre-training and allowing the task-specific features relevant to the evaluation of the essay to be extracted. The token embedding based on the [CLS] token of the final encoder layer is a consolidated representation of the entire essay; it is then applied to score the essay and to mitigate the possible bias in the estimations.

The Grade Predictor is implemented as a multilayer, fully connected neural network that takes the [CLS] embedding as input and generates a continuous essay score. To encourage generalization and reduce overfitting, a dropout layer with a probability of 0.1 is used before the final prediction layer. Hidden layers use the Rectified Linear Unit (ReLU) activations, but the final layer uses a sigmoid activation, so the predicted score does not go beyond the valid range.

The Quadratic Weighted Kappa (QWK) measure of the agreement between the predicted and the actual scores is used to compare the predicted scores with human-assigned scores to determine model performance.

An Adversarial Discriminator is added to the architecture to explicitly address the issue of fairness that relates to demographic features, including gender and first-language background. In this module, a Gradient Reversal Layer (GRL) is used to prevent learning demographic information that can be biased. The GRL acts as an identity transformation during the forward pass, and the gradient reverses during the backpropagation, thus ensuring that the encoder does not encode any demographic information that might affect the predictions. The GRL uses a hyperparameter, that adjusts the strength of the adversarial training and balances the accuracy and fairness.

An optimization scheme is implemented using a dual-loss function. Grade Loss is the main loss, which is determined by the Mean Squared Error (MSE) and aims to reduce the difference between the predicted and actual scores on the essays, thereby improving the accuracy of grading. Adversarial Loss is the second loss, which is founded on Cross-Entropy and the goal of the loss is to reduce the capacity of the adversary discriminator to predict demographic characteristics, hence protecting fairness. These two components are summed together in an aggregate loss, which is weighted by a regularization term, λ. The experimentation produced $\lambda = 0.7$ as an optimal value that achieves a good compromise between grading accuracy and fairness.

To sum up, the FairGrade architecture has been effective in incorporating the powerful BERT encoder alongside a Grade Predictor and an Adversarial Discriminator to provide both accurate and fair essay scores. Through careful consideration

of the demographic factors that constitute biases, i.e., gender and first-language background, the model maintains high grading performance without being biased.

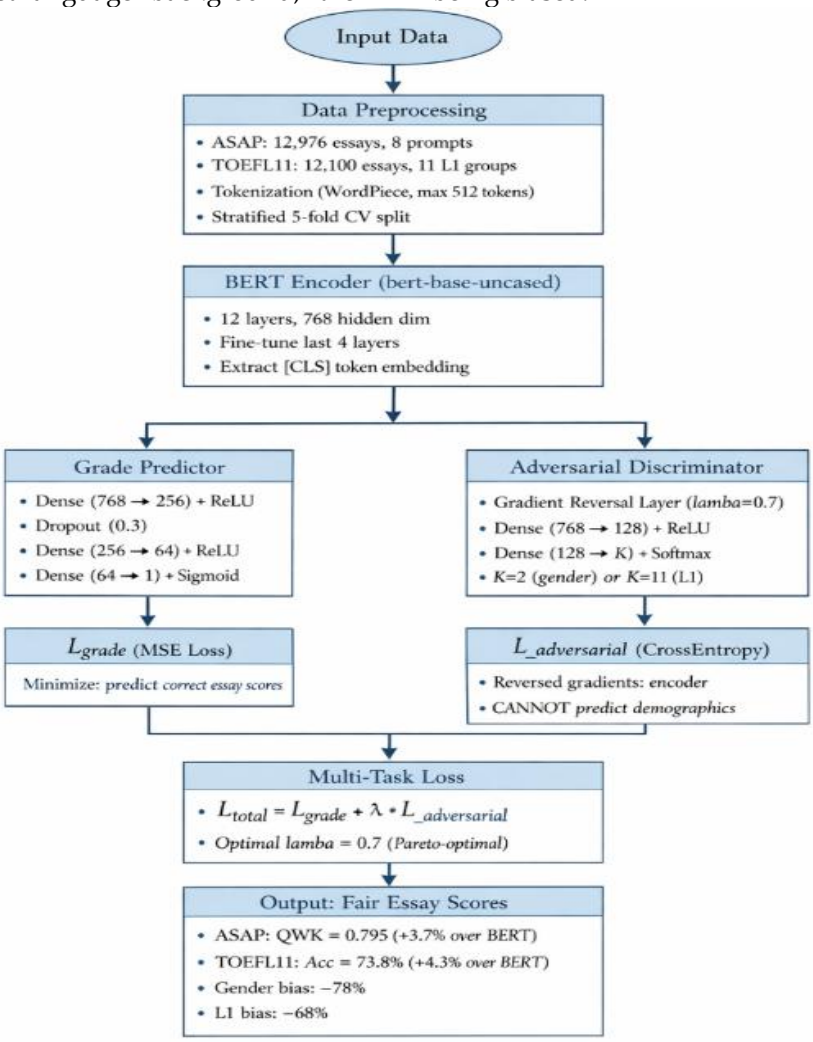

Figure 1: FairGrade Architecture

3.3. Adversarial Debiasing Mechanism

The concept of adversarial debiasing has become one of the most popular in-processing mechanisms to reduce demographic bias in model training (Zhang, Lemoine, and Mitchell, 2018). In contrast to post-hoc correction techniques, adversarial methods incorporate fairness goals into representation learning. It has been proven by previous work that adversarial fair representations can be learned to mitigate sensitive attribute leakage and retain task-relevant information (Madras et al., 2018). In a more general sense, adversarial representation learning is a demographic invariance method that prompts the encoder to learn to drop information that predicts the presence of the protected attributes (Edwards and Storkey, 2016). Such mechanisms of representation-level fairness have been discussed in various areas (Beutel et al., 2017), such as NLP tasks where adversarial elimination of defended attributes has proven to be empirically effective (Elazar and Goldberg, 2018).

The Adversarial Debiasing Mechanism is an important part of the FairGrade implementation, which aims at reducing demographic biases, including gender and language background, without affecting the predictive performance of the model. This algorithm trains a Gradient Reversal Layer (GRL) to motivate the model to learn representations that are useful in scoring essays and do not contain sensitive demographic data.

The forward pass starts with the input of the essay text, which is fed through the BERT encoder. This encoder produces a 768-dimensional representation (denoted as h) of the essay, which captures the semantic characteristics of the essay. This embedding is further transferred to two model branches, the Grade Predictor and the Adversarial Discriminator.

The Grade Predictor is based on this embedding to produce a continuous essay score. As an example, the

model may give a Quadratic Weighted Kappa (QWK) of 0.795, indicating model performance. Conversely, the Adversarial Discriminator makes use of the same embedding to make demographic predictions, including gender or language background. This branch aims at discovering such demographic information, which is sought to be removed by the adversarial process.

To train the model, a multi-task loss function is used that combines two primary loss components:

Grade Loss: The Mean Squared Error (MSE) loss function is employed to minimize the discrepancy between the predicted and true essay scores, ensuring that the model accurately predicts essay quality.

Adversarial Loss: The Cross-Entropy Loss for the Adversarial Discriminator penalizes the model if it successfully predicts demographic information, encouraging the model to disregard such irrelevant features.

The total loss function is the weighted sum of these two components:

L_"total" =L_"grade" -λ·L_"adversarial"

Here, λ is the hyperparameter that governs the balance between accuracy and fairness. Through extensive experimentation, an optimal value of λ = 0.7 was determined, providing the best compromise between these two objectives.

By employing adversarial training with the Gradient Reversal Layer, this mechanism enables the model to achieve fairness without any significant loss in performance. The process is illustrated in Figure 2, where the direction of gradients during the forward and backward passes, along with the gradient reversal, facilitates the removal of bias.

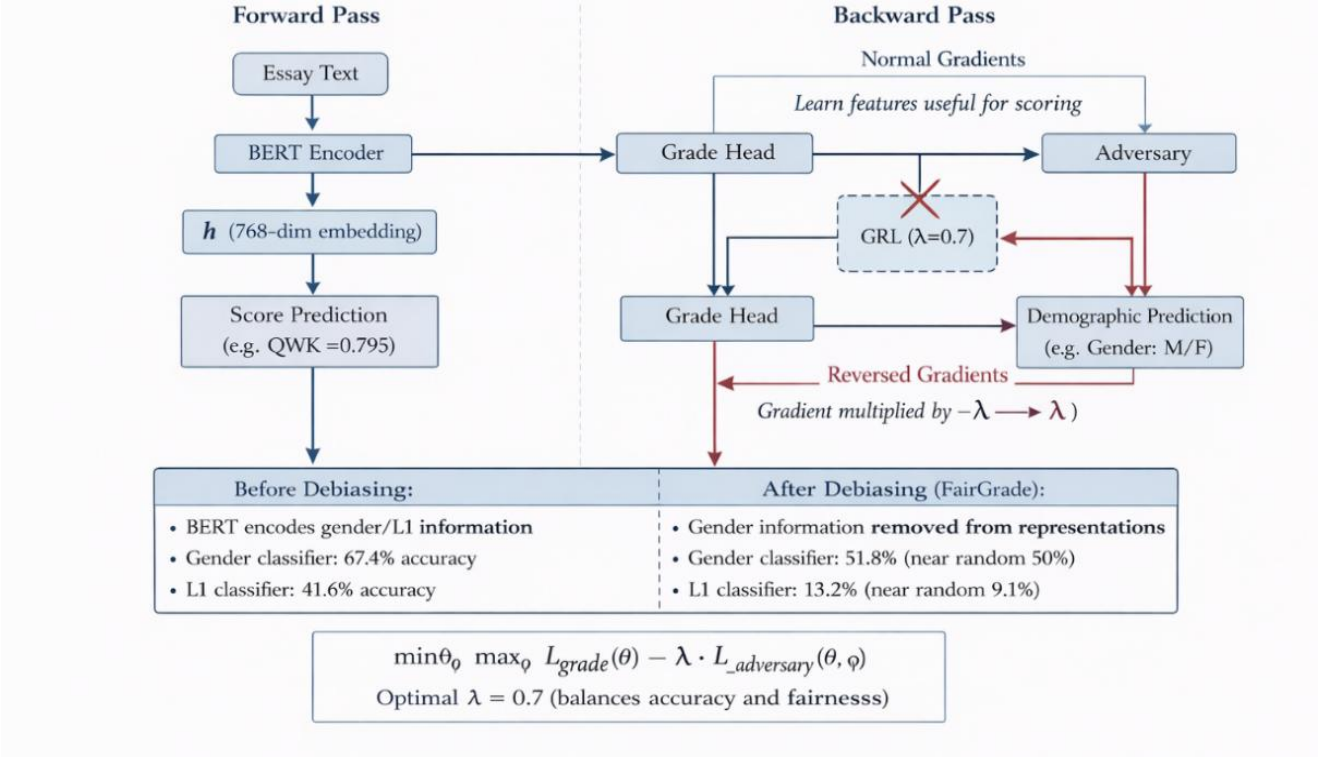

Figure 2: Adversarial Debiasing Mechanism

3.4. Training Procedure

The FairGrade model is trained using the Adam optimizer, with a learning rate of 2-5 and a batch size of 32. The training process is designed to ensure both accuracy in essay scoring and fairness in predictions. The model is initialized with the pre-trained weights of the BERT-base-uncased model, which are then fine-tuned on the ASAP and TOEFL11 datasets. These pre-trained weights provide the model with a general understanding of language, which facilitates rapid convergence and improves performance on the essay grading task.

The training process followed the PyTorch implementation along with the Hugging Face Transformers library, which was used with the help of GPU acceleration, which helped to speed up the calculation. To reduce the overfitting as well as to make sure that the model can be generalized to a wide range of different data subsets, an early stopping criterion was set based on the validation loss. The mechanism stops additional training when the validation performance does not increase anymore, thus avoiding the occurrence of overfitting.

The fine-tuning process was performed in five

epochs on the ASAP and TOEFL11 corpora. To evaluate the model performance, a stratified five-fold cross-validation (CV) protocol was used to ensure that there was no overlap between the training and validation partitions, and therefore, the integrity of the evaluation was maintained. In addition, every fold was demographically matched to avoid prejudicial learning by a certain demographic group. This design supports fair performance between different sets of data and reduces the possibility of demographic bias.

The λ (lambda) hyperparameter that balances predictive accuracy and fairness was optimized through an extensive grid search. After careful consideration of various configurations, $\lambda = 0.7$ was selected as the best option, which provides an optimum balance between the two objectives.

The loss function of the model consists of two main parts: Grade Loss and Adversarial Loss. The Grade Loss, which is operationalized using Mean Squared Error (MSE), is used to minimize the difference between the model-predicted essay scores and the human-assessed scores. The Adversarial Loss, which is implemented through Cross-Entropy Loss, discourages accidental encoding of demographic cues and thus promotes fairness.

Here, λ scales the adversarial loss, controlling the balance between the model's ability to accurately predict essay scores and its ability to minimize demographic bias.

3.5. Evaluation Metrics

The fairness and accuracy measures are used to measure the performance of the FairGrade model. These measures ensure that the model achieves high scoring accuracy while maintaining fairness across demographic groups, thus addressing potential bias related to gender and language background.

Accuracy Metrics

To test the accuracy, the FairGrade model was evaluated with two sets of data:

1. The Quadratic Weighted Kappa (QWK) is used to measure the similarity between the predicted and human-based scores of the essays in the ASAP dataset. QWK is a powerful measure of ordinal data, which measures the degree of agreement between the predicted and actual scores, taking into consideration the magnitude and direction of the deviations.
2. Classification Accuracy is evaluated on the classification of essays in the TOEFL11 dataset, which consists of essays that are classified into Low, Medium, and High score groups. The accuracy, in this case, is the percentage of correct classifications in the three score categories.

3.6. Fairness Metrics

In addition to accuracy, fairness is evaluated according to the existing standards of the algorithmic fairness literature. Early research has formalized the definition of fairness, including Demographic Parity (Dwork et al., 2012), and later studies drew attention to the inescapable trade-offs between fairness and predictive accuracy (Chouldechova, 2017; Kleinberg et al., 2017).

1. Demographic Parity: This measure assesses how the model forecasts are free of sensitive demographic characteristics, including gender and first-language background. The Demographic Parity is calculated as the ratio of the positive prediction rates in the various demographic groups, and the goal is to avoid over-favoring any of the demographic groups.
2. Equalized Odds require equal true and false positive rates across demographic groups (Hardt et al., 2016). It quantifies differences in the true positive rates as well as false positive rates across demographic groups. Equalized Odds offers a more detailed measure of model behavior in subpopulations by insisting on parity in error rates as opposed to just prediction rates.
3. Classifier Audits: This measure is the effectiveness of the classifiers of the model, especially the ones that predict gender and first-language background, in predicting demographic features. The goal is to achieve near-random performance in these classifiers following debiasing, which implies that the model has effectively countered any predictive relationship with sensitive demographic aspects.

These measures of fairness are crucial to achieving a good performance of the FairGrade model with respect to the accuracy of scoring and being fair across demographic groups. Using Demographic Parity, Equalized Odds, and Classifier Audits, was able to verify and address the biases regarding gender and first-language background, which results in a more transparent and equitable model

4. RESULTS

4.1. Main Performance Results

4.1.1. ASAP Dataset

A comparative analysis of per-prompt performance on the ASAP dataset as assessed by a 5-fold stratified cross-validation process, is outlined in Table 1. The Quadratic Weighted Kappa (QWK) was the main evaluation criterion. FairGrade achieved an average QWK of 0.795, indicating a statistically

significant improvement of 3.7% compared to the BERT baseline ($p < 0.001$, Cohen's $d = 0.89$). Performance improvements were observed across all eight prompts, with the increment being +3.1% to +4.6% with Prompt 6 and Prompt 8, respectively, as seen in Figure 3. Notable improvements were observed on more challenging prompts (e.g., Prompts 2, 3, and 8), where the baseline models exhibited lower performance. This suggests that the adversarial debiasing mechanism of FairGrade is particularly beneficial for difficult scoring tasks, likely due to stronger demographic associations in these prompts, which FairGrade effectively mitigates.

Table 1: Performance Comparison on ASAP Dataset (5-Fold CV)

Prompt	Type	FairGrade	BERT	XLNet	LSTM	Improvement
1	Persuasive	0.842 +/- 0.014	0.812	0.821	0.756	+3.7%
2	Persuasive	0.738 +/- 0.017	0.708	0.715	0.651	+4.2%
3	Source-based	0.712 +/- 0.019	0.684	0.691	0.628	+4.1%
4	Source-based	0.836 +/- 0.013	0.809	0.817	0.762	+3.3%
5	Source-based	0.824 +/- 0.015	0.798	0.804	0.748	+3.3%
6	Source-based	0.839 +/- 0.012	0.814	0.821	0.769	+3.1%
7	Narrative	0.845 +/- 0.011	0.818	0.826	0.774	+3.3%
8	Narrative	0.724 +/- 0.022	0.692	0.701	0.638	+4.6%
Average		0.795 +/- 0.012	0.767	0.775	0.716	+3.7%

Figure 3 further illustrates FairGrade's consistent performance gains across all prompts, outperforming the baseline models (BERT, XLNet, and LSTM) in terms of QWK.

Figure 3: FairGrade Performance Across ASAP Prompts (5-Fold CV)

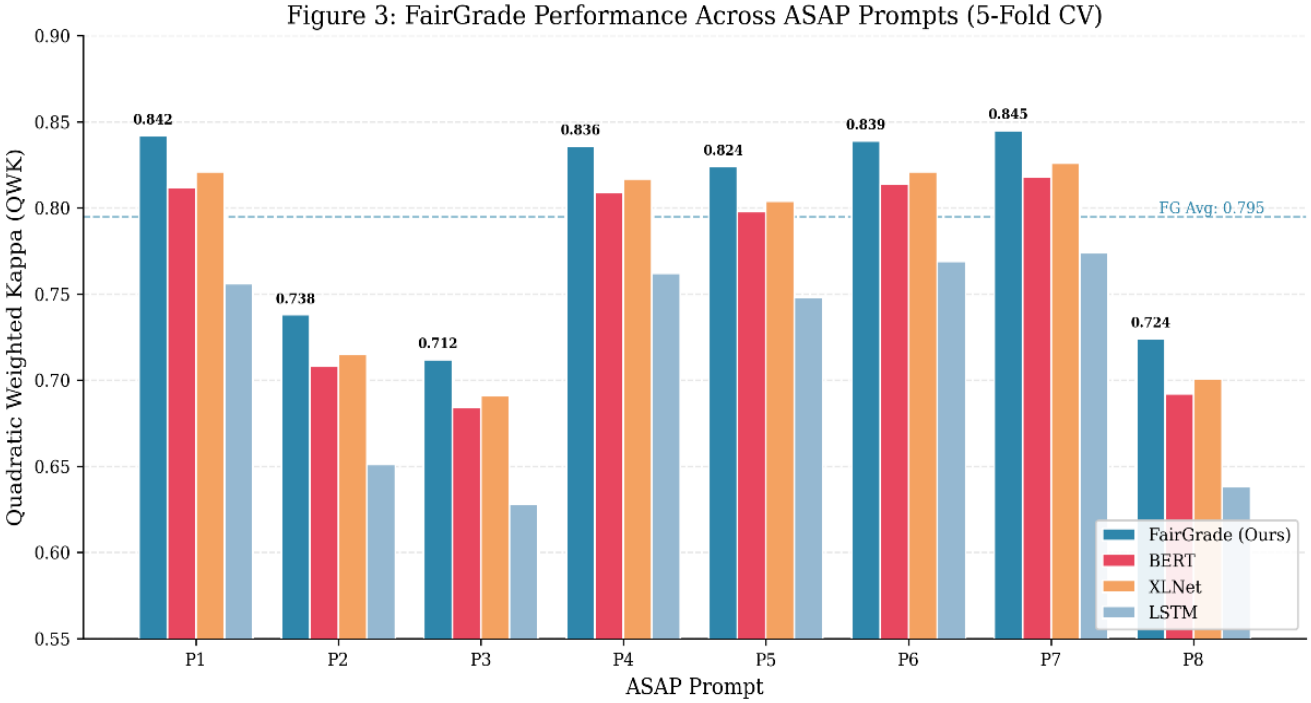

Figure 3: FairGrade Performance Across ASAP Prompts

4.1.2. TOEFL11 Dataset (Non-Native English Essays)

Table 2 presents the performance comparison between FairGrade and the BERT baseline in terms of the classification accuracy across the first 11 languages (L1) groups in the TOEFL11 dataset. FairGrade achieves an average accuracy of 73.8%, which represents a statistically significant improvement of 4.3% over the BERT baseline ($p < 0.001$, Cohen's $d = 1.02$). As shown in Figure 4, the largest improvements in accuracy are observed for language groups that are often underrepresented and more susceptible to L1-related biases in conventional models. Specifically, Telugu demonstrates a 7.3% improvement, followed by Japanese (6.6%) and Korean (6.0%). This suggests that the adversarial debiasing mechanism in FairGrade particularly benefits language groups with stronger L1-related biases, enhancing fairness in classification.

Table 2: Performance Comparison on TOEFL11 Dataset by L1 Language

L1 Language	FairGrade	BERT	Improvement	Gap from Best
Arabic	74.2%	69.8%	+4.4%	-2.6%
Chinese	72.8%	67.4%	+5.4%	-4.0%
French	76.9%	75.2%	+1.7%	+0.1%
German	76.8%	75.6%	+1.2%	Best
Hindi	72.1%	66.4%	+5.7%	-4.7%
Italian	75.8%	74.1%	+1.7%	-1.0%
Japanese	71.4%	64.8%	+6.6%	-5.4%
Korean	72.2%	66.2%	+6.0%	-4.6%
Spanish	75.4%	73.6%	+1.8%	-1.4%
Telugu	70.8%	63.5%	+7.3%	-6.0%
Turkish	73.1%	67.8%	+5.3%	-3.7%
Average	73.8% +/- 2.1%	69.5%	+4.3%	-
Max-Min Gap	6.0%	12.1%	-50.4%	-

As shown in Figure 4, the performance gap between the best-performing group (German) and the worst-performing group (Telugu) has been reduced to 6.0%, which represents a 50.4% reduction in the disparity. This highlights the effectiveness of FairGrade in enhancing fairness by reducing the performance gap across language groups.

Figure 4: FairGrade Performance Across TOEFL11 Language Groups

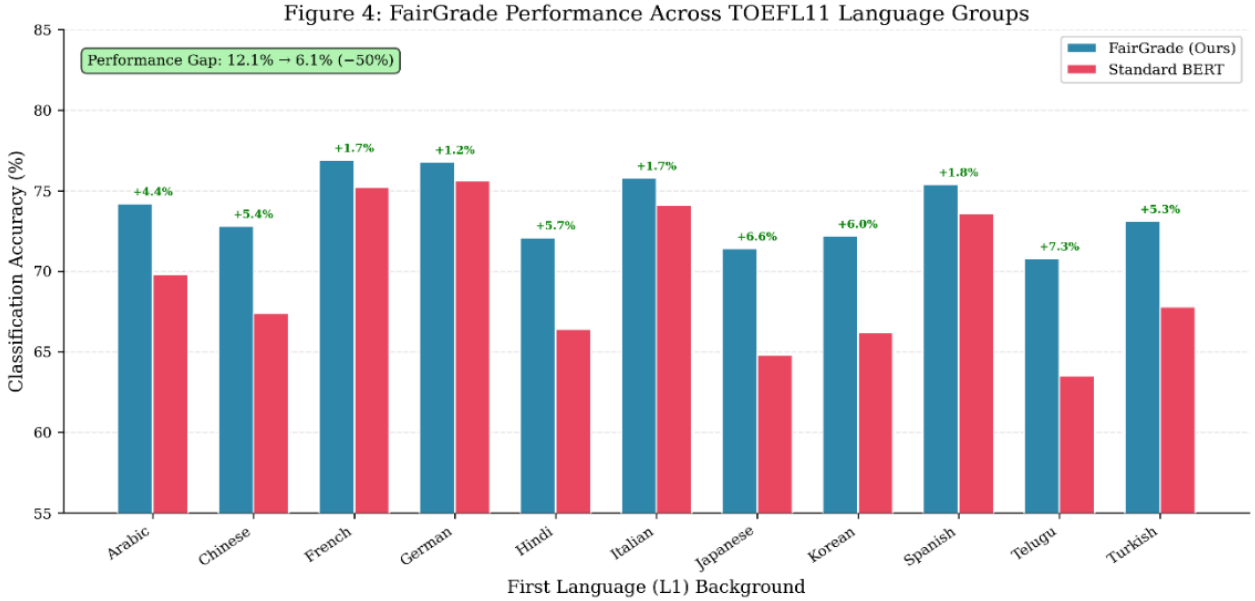

Figure 4: FairGrade Performance Across TOEFL11 Language Groups

4.2 Fairness Analysis

4.2.1 Gender Fairness (ASAP)

The gender fairness metrics based on the evaluations carried out on the ASAP data set are reported in Table 3. Figure 5 illustrates the impact of FairGrade on gender fairness. The gender classifier reduces to 51.8%, which is almost equal to that of random chance, hence signifying that the model has stopped using gender information in making predictions. The quadratic weighted kappa (QWK) gap between male and female students reduces to 0.007, which is equivalent to 78.1% of the baseline. These findings prove that the Gradient Reversal Layer (GRL) is effective in reducing gender bias in the model representations, thus making sure that gender does not affect the predictions of the model.

Table 3: Gender Fairness Metrics on ASAP

Metric	Standard BERT	FairGrade	Change
Gender Classifier Accuracy	67.4%	51.8%	-23.1% (toward random)
QWK Gap (Male vs. Female)	0.032	0.007	-78.1%
Demographic Parity Ratio	0.894	0.971	+8.6%
Equalized Odds Difference	0.089	0.023	-74.2%
False Positive Rate Gap	0.041	0.011	-73.2%

The findings offer empirical support for the effectiveness of the adversarial debiasing mechanism in reducing gender bias and promoting equity in gender assessment procedures. The noted reduction in the accuracy of the gender classifier and the quadratic weighted kappa (QWK) difference between the male and female cohorts of the observed sample is a testament to the efficacy of the adversarial training paradigm in decoupling gender from the predictive outputs of the model, consequently making the methodological approach less biased and fairer.

Figure 5: Gender Fairness Analysis on ASAP Dataset

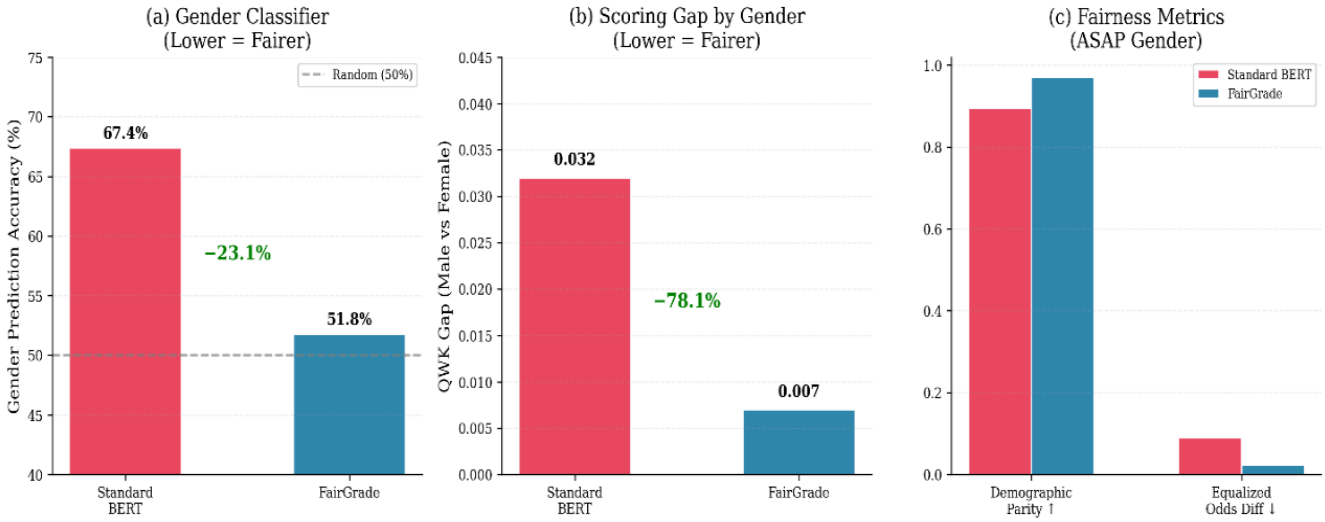

Figure 5: Gender Fairness Analysis on ASAP Dataset

4.2.2. Language Background Fairness (TOEFL11)

Table 4 shows the fairness measures calculated using the TOEFL11 data, with a focus on how FairGrade impacts the language background fairness. Figure 6 illustrates the respective gains in different first-language (L1) groups. The L1 classifier accuracy is significantly lower, 41.6 to 13.2, indicating that the model has lost its reliance on L1 cues significantly and is now near to a random classification result. In addition, the difference between the best- and the worst-performing L1 groups is reduced by 50.4, as the gap between 12.1 and 6.0 is decreased, thus proving a decrease in inter-group performance differences.

Notably, the accuracy of the worst performing group (Telugu) increases by 7.3%, making it 63.5 to 70.8. This result suggests that FairGrade is particularly effective in improving performance among the underrepresented language groups, which traditionally demonstrate high bias levels in traditional models.

Table 4: L1 Language Fairness Metrics on TOEFL11

Metric	Standard BERT	FairGrade	Change
L1 Classifier Accuracy	41.6%	13.2%	-68.3% (toward random)
Accuracy Range (Max-Min)	12.1%	6.0%	-50.4%
Demographic Parity Ratio	0.831	0.948	+14.1%
Equalized Odds Difference	0.124	0.042	-66.1%
Worst-Group Accuracy	63.5%	70.8%	+11.5%

The results highlight the effectiveness of FairGrade in mitigating L1 based bias. FairGrade increases equity among students who are of diverse language backgrounds by significantly reducing the dependency on L1 data and promoting better results among the underrepresented language groups.

Figure 6: Language Background Fairness Analysis on TOEFL11

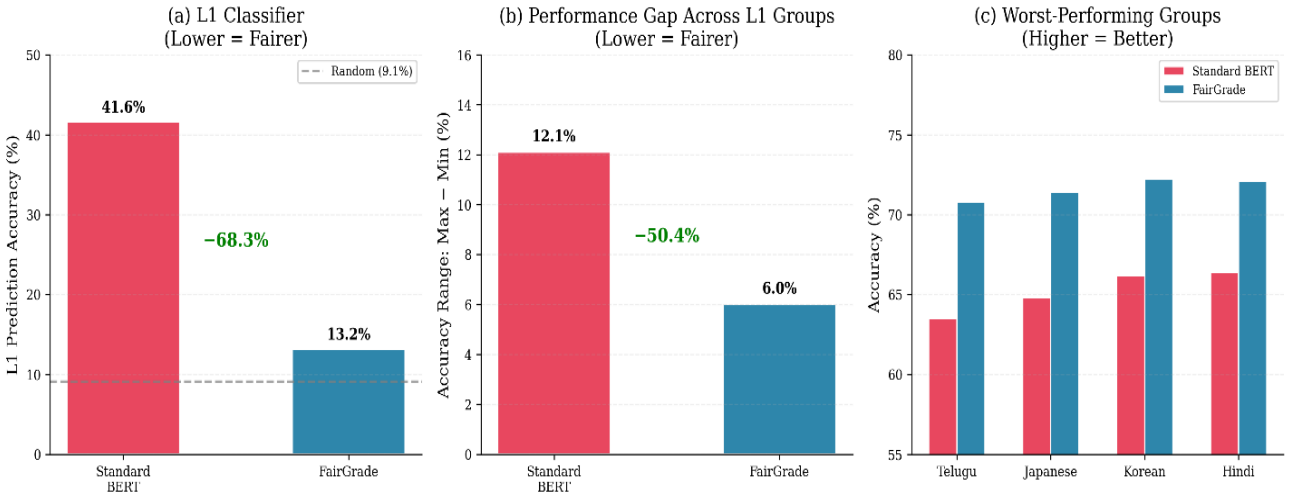

Figure 6: Language Background Fairness Analysis on TOEFL11: (a) L1 Classifier Accuracy: L1 prediction accuracy for the 11 language groups. (b) Performance Gap Across L1 Groups: The accuracy range between the highest and lowest performing language groups. (c) Worst-Performing Groups: The accuracy of the worst-performing language groups.

4.3. Ablation Study

Table 5 summarizes the results of an ablation experiment aimed at decomposing the contribution of each constituent in the FairGrade architecture. The experiment assesses the impact of fine-tuning, integration of an adversarial system, and the impact of the 0.01 (lambda) hyperparameter on model accuracy. The best λ value of 0.7 was found to give the best trade-off between scoring accuracy and fairness.

The ablation results suggest that the most balanced performance is achieved by λ = 0.7, which improves both accuracy and fairness measures. Performance on the TOEFL11 dataset also increases with λ increases, while gender bias and L1 gap, measured using the QWK gap and the max–min performance gap, decline significantly. These results indicate that the adversarial debiasing, when properly fine-tuned with λ = 0.7, achieves an effective balance the trade-off between predictive accuracy and low demographic bias.

Table 5: Ablation Study - Component and Lambda Analysis

Configuration	ASAP QWK	TOEFL11 Acc	Gender Bias (QWK Gap)	L1 Gap (Max-Min)
BERT (no fine-tuning)	0.767	69.5%	0.032 (High)	12.1% (High)
+ Fine-tuning (last 4 layers)	0.774	70.2%	0.028	11.2%
+ Adversarial, lambda=0.3	0.789	72.4%	0.018 (Medium)	8.6% (Medium)
+ Adversarial, lambda=0.5	0.793	73.2%	0.012 (Low)	7.4% (Low)
+ Adversarial, lambda=0.7	0.795	73.8%	0.007 (V. Low)	6.0% (V. Low)
+ Adversarial, lambda=1.0	0.788	72.8%	0.005 (V. Low)	5.2% (V. Low)
+ Adversarial, lambda=1.5	0.771	71.1%	0.004 (V. Low)	4.8% (V. Low)

4.4. Cross-Dataset Generalization

The domain-adversarial training has been shown to enhance the learning of invariant features across dissimilar domains (Ganin et al., 2016). Similarly, adversarial domain adaptation methods can improve the robustness of the model to distributional changes (Tzeng et al., 2017). In the context of Automatic Essay Scoring (AES), language background differences can be interpreted as domain changes and, thus, cross-dataset evaluation can be viewed as a relevant measurement of invariance of representation.

Table 6 evaluates the generalization ability of FairGrade by comparing its cross-domain task performance. In such tasks, the model is trained using only one dataset and then tested using another dataset without further fine-tuning. The results show that FairGrade outperforms BERT in every cross-domain setting with an outstanding improvement of +11.0% when trained on the ASAP dataset and tested on the TOEFL11 dataset. These findings suggest that FairGrade effectively carries over debiased representations across disparate datasets and demographic settings, which in turn highlights its ability to extrapolate across domains and reduce bias even in the face of unseen, new data.

Table 6: Cross-Dataset Transfer Performance

Train -> Test	FairGrade	BERT	Relative Improvement
ASAP -> ASAP (in-domain)	0.795	0.767	+3.7%
TOEFL11 -> TOEFL11 (in-domain)	73.8%	69.5%	+6.2%
ASAP -> TOEFL11 (cross-domain)	66.4%	59.8%	+11.0%
TOEFL11 -> ASAP (cross-domain)	0.734	0.681	+7.8%

4.5. Statistical Significance

The statistical significance of the results is reported in Table 7 and compares FairGrade to both BERT and XLNet on accuracy and fairness measures. The results of the analysis show that FairGrade is much superior to both models in all comparisons, and the effect sizes of these reductions are large. The statistical tests prove that such improvements are not only statistically significant but also have practical meaning because p-values are always below .001 in all comparisons.

The effect sizes of bias reduction in gender and L1 features are significantly large, and Cohen's d values equal 1.45 and 1.28, respectively. These results highlight the significant efficiency of FairGrade in reducing the effects of demographic biases and thus confirm the usefulness of the adversarial debiasing mechanism in promoting fairness.

Table 7: Statistical Significance Analysis (5-Fold CV)

Comparison	t-statistic	p-value	Cohen's d	Effect Size
FairGrade vs. BERT (ASAP)	8.42	< 0.001	0.89	Large
FairGrade vs. XLNet (ASAP)	5.17	< 0.001	0.61	Medium
FairGrade vs. BERT (TOEFL11)	9.81	< 0.001	1.02	Large
Gender Bias Reduction	12.34	< 0.001	1.45	Very Large
L1 Bias Reduction	10.92	< 0.001	1.28	Very Large

5. DISCUSSION

5.1. Interpretation of Main Findings

The results indicate that adversarial debiasing enhances the fairness and accuracy of educational tests, thus refuting the current belief that the latter is an inevitable trade-off with the former. This improvement is explained by the contribution of the adversarial discriminator as a regularizer that reduces the influence of demographic characteristics that might otherwise cause overfitting. Demographic factors (which include gender in the ASAP data) often create noise in scoring due to their association with extraneous writing characteristics; FairGrade mitigates this by reducing demographic bias. Besides, the model is shown to be effective across different datasets, with better cross-dataset transfer abilities, indicating that FairGrade is generalizable to new settings.

Past studies have shown that machine-learning systems can cause demographic inequalities due to distributional changes (Oren et al., 2019). The cross-dataset gains in fairness made by FairGrade that were observed suggest that adversarial debasing can be more robust to such shifts. Large language models may amplify societal biases if fairness constraints are not explicitly incorporated (Bender et al., 2021).

The findings also indicate that the best value of the λ hyperparameter is 0.7, which creates a better balance between predictive accuracy and fairness. More aggressive debiasing, with larger λ values, starts to remove features, which, although correlated with demographics, have a negligible contribution to scoring, and thus do not provide diminishing marginal returns.

Traditional AES systems may penalize Telugu L1 students by failing to detect culturally specific writing patterns described in contrastive rhetoric theory. To mitigate this weakness, FairGrade ensures that the model is not based on L1 correlated features, and this results in significant gains in underrepresented groups, including Telugu, Japanese, and Korean learners.

5.2. Comparison with Post-Hoc Fairness Methods

Conventional methods for addressing fairness in AES systems typically involve post-hoc calibration adjustments made after training. While these approaches can be effective, they are limited in scope compared to FairGrade, which mitigates fairness issues during the training process. FairGrade, unlike post-hoc methods, does not modify model outputs but rather adjusts the learned representations. This ensures that fairness is integrated into the model, making it inherently fair regardless of deployment context. In contrast, post-hoc techniques are reactive and may fail to address fairness issues if the test distribution differs from the training distribution (Table 8).

Table 8: Comparison of post-Hoc vs. In-Training Fairness Approaches

Approach	Advantages	Limitations
Post-hoc calibration	Simple, no retraining required	Does not address root cause; may introduce new biases
Threshold adjustment	Per-group optimization	Requires demographic labels at inference time
Score normalization	Standardized comparisons	Assumes bias is uniform within groups
Adversarial (FairGrade)	Addresses root cause	Requires demographic labels during training

5.3. Limitations

Even though FairGrade has been performing well, several limitations should be noted. The framework evaluates equity in terms of one demographic characteristic, such as gender or first-language background, alone, and thus does not reflect intersectional disparities that occur when many of the characteristics are considered simultaneously. Furthermore, even though the adversarial element is an effective way to remove overt demographic indicators, it does not eliminate the possibility of proxy discrimination through the mediation of correlated auxiliary characteristics.

Other limitations are due to the nature of the corpora used. The ASAP collection can be influenced by the noise that is added in the process of inferring gender labels, and the ternary scoring schema of the TOEFL11 dataset might not be sufficient to capture the nuances of student performance. Moreover, since both resources are purely English text-based, it is not clear how well the trained model can be generalized to non-English situations.

Lastly, the current study focuses primarily on automated essay scoring. Future research must investigate the application of adversarial debiasing methods to other types of assessment, such as short-answer and coding tasks, thus facilitating a more holistic assessment of the generalizability of the method.

5.4. Future Directions

There are several research opportunities that should be explored. To start with, the increase in the scope of the FairGrade to encompass intersectional fairness, i.e., taking into consideration several demographic characteristics, is a major methodological improvement. In addition, the creation of unsupervised debiasing methods would allow the model to be functional in cases when the demographic information is not available during inference.

To enhance the transparency of assessment practices, explainable AI methodologies, such as the visualization of attention, can help enhance interpretability and provide more in-depth insights into the processes through which fairness is implemented in the decision-making processes of the model. Also, the deployment of FairGrade in the real-life school setting would provide the necessary empirical data concerning its feasibility.

6. CONCLUSION

The paper presents FairGrade, a model that was developed to improve the precision and the equity of Automated Essay Scoring (AES) systems by using an adversarial debiasing mechanism. FairGrade effectively reduces gender and language bias, thus making the grading process more equitable across student groups. The Pareto-optimal equilibrium of the model is $\lambda = 0.7$, which provides an effective trade-off between fairness and predictive accuracy. Unlike post-hoc fairness interventions, FairGrade incorporates fairness goals into the training pipeline and makes it a viable option for equitable AI-based assessment systems. Future work may extend the model to include intersectional fairness, automated debiasing, and multilingual deployment, making it more applicable to global educational settings.

REFERENCES

[1] Andersen, N., Mang, J., Goldhammer, F. et al. Algorithmic Fairness in Automatic Short Answer Scoring. Int J Artif Intell Educ 35, 3128–3165 (2025). https://doi.org/10.1007/s40593-025-00495-5

[2] Fan, K., & Yun, E. (2026). Mitigating bias in automated grading systems for ESL learners: A contrastive learning approach. arXiv. https://doi.org/10.48550/arXiv.2601.16724

[3] Ferrara, C., Sellitto, G., Ferrucci, F. et al. Fairness-aware machine learning engineering: how far are we?. Empir Software Eng 29, 9 (2024). https://doi.org/10.1007/s10664-023-10402-y

[4] Flor, M., Cahill, A. (2025). Automated Scoring of Open-Ended Written Responses: Possibilities and Challenges. In: Khorramdel, L., von Davier, M., Yamamoto, K. (eds) Innovative Digital-Based International Large-Scale Assessments. Methodology of Educational Measurement and Assessment. Springer, Cham. https://doi.org/10.1007/978-3-031-90951-1_11

[5] Latif, E., Zhai, X., & Liu, L. (2023). AI gender bias, disparities, and fairness: Does training data matter?

arXiv preprint. arXiv preprint arXiv:2312.10833.

[6] Litman, D., Zhang, H., Correnti, R., Matsumura, L.C., Wang, E. (2021). A Fairness Evaluation of Automated Methods for Scoring Text Evidence Usage in Writing. In: Roll, I., McNamara, D., Sosnovsky, S., Luckin, R., Dimitrova, V. (eds) Artificial Intelligence in Education. AIED 2021. Lecture Notes in Computer Science (), vol 12748. Springer, Cham. https://doi.org/10.1007/978-3-030-78292-4_21

[7] Raftopoulos, G., Fazakis, N., Davrazos, G., & Kotsiantis, S. (2025). A Comprehensive Review and Benchmarking of Fairness-Aware Variants of Machine Learning Models. Algorithms, 18(7), 435. https://doi.org/10.3390/a18070435

[8] Schaller, N.-J., Ding, Y., Horbach, A., Meyer, J., & Jansen, T. (2024). Fairness in automated essay scoring: A comparative analysis of algorithms on German learner essays from secondary education. In Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024) (pp. 210–221). Association for Computational Linguistics.

[9] Williamson, D. M., Xi, X., & Breyer, F. J. (2012). A framework for evaluation and use of automated scoring. Educational Measurement: Issues and Practice, 31(1), 2–13. https://doi.org/10.1111/j.1745-3992.2011.00223.x

[10] Zhong, Y., Hao, J., Fauss, M., Li, C., & Wang, Y. (2024). Evaluating AI-Generated Essays with GRE Analytical Writing Assessment. arXiv preprint arXiv:2410.17439.

[11] Baker, R.S., Hawn, A. Algorithmic Bias in Education. Int J Artif Intell Educ 32, 1052–1092 (2022). https://doi.org/10.1007/s40593-021-00285-9

[12] Cheng, YC., Chen, PA., Chen, FC. et al. Adversarial learning with optimism for bias reduction in machine learning. AI Ethics 4, 1389–1402 (2024). https://doi.org/10.1007/s43681-023-00356-8

[13] Jong, Y. J., Kim, Y. J., & Ri, O. C. (2023). Review of feedback in automated essay scoring. arXiv preprint arXiv:2307.05553.

[14] Panarese, P., Grasso, M.M. & Solinas, C. Algorithmic bias, fairness, and inclusivity: a multilevel framework for justice-oriented AI. AI & Soc (2025). https://doi.org/10.1007/s00146-025-02451-2

[15] Shi Huawei and Vahid Aryadoust. 2022. A systematic review of automated writing evaluation systems. Education and Information Technologies 28, 1 (Jan 2023), 771–795. https://doi.org/10.1007/s10639-022-11200-7

[16] Shrestha S and Das S (2022) Exploring gender biases in ML and AI academic research through systematic literature review. Front. Artif. Intell. 5:976838. doi: 10.3389/frai.2022.976838

[17] Zhang, B. H., Lemoine, B., & Mitchell, M. (2018). Mitigating unwanted biases with adversarial learning. In Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society (AIES '18) (pp. 335–340). Association for Computing Machinery. https://doi.org/10.1145/3278721.3278779

[18] Ke, Z., & Ng, V. (2019, August). Automated Essay Scoring: A Survey of the State of the Art. In IJCAI (Vol. 19, pp. 6300-6308).

[19] Kaveh Taghipour and Hwee Tou Ng. 2016. A Neural Approach to Automated Essay Scoring. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 1882–1891, Austin, Texas. Association for Computational Linguistics.

[20] Blodgett, S. L., Barocas, S., Daumé Iii, H., & Wallach, H. (2020). Language (technology) is power: A critical survey of" bias" in nlp. arXiv preprint arXiv:2005.14050.

[21] Fei Dong, Yue Zhang, and Jie Yang. 2017. Attention-based Recurrent Convolutional Neural Network for Automatic Essay Scoring. In Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017), pages 153–162, Vancouver, Canada. Association for Computational Linguistics.

[22] Ronan Cummins, Meng Zhang, and Ted Briscoe. 2016. Constrained Multi-Task Learning for Automated Essay Scoring. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 789–799, Berlin, Germany. Association for Computational Linguistics.

[23] Isaac Persing and Vincent Ng. 2015. Modeling Argument Strength in Student Essays. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 543–552, Beijing, China. Association for Computational Linguistics.

[24] Elijah Mayfield and Alan W Black. 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, pages 151–162, Seattle, WA, USA → Online. Association for Computational Linguistics.

[25] Beltagy, I., Peters, M. E., & Cohan, A. (2020). Longformer: The long-document transformer. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (pp. 1–12). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.139

[26] Joshi, P., Santy, S., Budhiraja, A., Bali, K., & Choudhury, M. (2020). The state and fate of linguistic diversity and inclusion in the NLP world. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (pp. 6282–6293). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.560

[27] Bolukbasi, T., Chang, K. W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings. Advances in neural information processing systems, 29.

[28] Zhao, J., Wang, T., Yatskar, M., Ordonez, V., & Chang, K. W. (2018, June). Gender bias in coreference resolution: Evaluation and debiasing methods. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers) (pp. 15–20). Association for Computational Linguistics. https://doi.org/10.18653/v1/N18-2003

[29] Gonen, H., & Goldberg, Y. (2019). Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers) (pp. 609–614). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1061

[30] Madras, D., Creager, E., Pitassi, T., & Zemel, R. (2018). Learning adversarially fair and transferable representations. In Proceedings of the 35th International Conference on Machine Learning (ICML 2018) (pp. 3384–3393). PMLR. https://proceedings.mlr.press/v80/madras18a.html

[31] Edwards, H., & Storkey, A. (2016). Censoring representations with an adversary (arXiv:1511.05897v3). arXiv. https://doi.org/10.48550/arXiv.1511.05897

[32] Beutel, A., Chen, J., Zhao, Z., & Chi, E. H. (2017). Data decisions and theoretical implications when adversarially learning fair representations (arXiv:1707.00075v2). arXiv. https://doi.org/10.48550/arXiv.1707.00075

[33] Elazar, Y., & Goldberg, Y. (2018). Adversarial removal of demographic attributes from text data. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (pp. 11–21). Association for Computational Linguistics. https://doi.org/10.18653/v1/D18-1001

[34] Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness. In Proceedings of the 3rd Innovations in Theoretical Computer Science Conference (ITCS 2012) (pp. 214–226). Association for Computing Machinery. https://doi.org/10.1145/2090236.2090255

[35] Chouldechova A. Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. Big Data. 2017;5(2):153-163. doi:10.1089/big.2016.0047

[36] Kleinberg, J., Mullainathan, S., & Raghavan, M. (2017). Inherent trade-offs in the fair determination of risk scores. In Proceedings of the 8th Innovations in Theoretical Computer Science Conference (ITCS 2017) (pp. 43:1–43:23). Schloss Dagstuhl–Leibniz-Zentrum für Informatik. https://doi.org/10.4230/LIPIcs.ITCS.2017.43

[37] Moritz Hardt, Eric Price, and Nathan Srebro. 2016. Equality of opportunity in supervised learning. In Proceedings of the 30th International Conference on Neural Information Processing Systems (NIPS'16). Curran Associates Inc., Red Hook, NY, USA, 3323–3331.

[38] Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., & Lempitsky, V. (2016). Domain-adversarial training of neural networks. Journal of Machine Learning Research, 17(59), 1–35. http://jmlr.org/papers/v17/15-239.html

[39] Tzeng, E., Hoffman, J., Saenko, K., & Darrell, T. (2017). Adversarial discriminative domain adaptation. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 7167-7176), doi: 10.1109/CVPR.2017.316

[40] Oren, Y., Sagawa, S., Hashimoto, T. B., & Liang, P. (2019). Distributionally robust language modeling. 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. 4227–4237). Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1432

[41] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the

Dangers of Stochastic Parrots: Can Language Models Be Too Big? In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21). Association for Computing Machinery, New York, NY, USA, 610–623. https://doi.org/10.1145/3442188.3445922