Unveiling the Tapestry of Automated Essay Scoring: A Comprehensive Investigation of Accuracy, Fairness, and Generalizability

Kaixun Yang[1], Mladen Raković[1], Yuyang Li[1], Quanlong Guan[2*], Dragan Gašević[1], Guanliang Chen[1*]

[1]Centre for Learning Analytics at Monash, Monash University, Australia
[2]Department of Computer Science, Jinan University, China
{Kaixun.Yang1, Mladen.Rakovic, Dragan.Gasevic, Guanliang.Chen}@monash.edu, gql@jnu.edu.cn

Abstract

Automatic Essay Scoring (AES) is a well-established educational pursuit that employs machine learning to evaluate student-authored essays. While much effort has been made in this area, current research primarily focuses on either (i) boosting the predictive accuracy of an AES model for a specific prompt (i.e., developing prompt-specific models), which often heavily relies on the use of the labeled data from the same target prompt; or (ii) assessing the applicability of AES models developed on non-target prompts to the intended target prompt (i.e., developing the AES models in a cross-prompt setting). Given the inherent bias in machine learning and its potential impact on marginalized groups, it is imperative to investigate whether such bias exists in current AES methods and, if identified, how it intervenes with an AES model's accuracy and generalizability. Thus, our study aimed to uncover the intricate relationship between an AES model's accuracy, fairness, and generalizability, contributing practical insights for developing effective AES models in real-world education. To this end, we meticulously selected nine prominent AES methods and evaluated their performance using seven distinct metrics on an open-sourced dataset, which contains over 25,000 essays and various demographic information about students such as gender, English language learner status, and economic status. Through extensive evaluations, we demonstrated that: (1) prompt-specific models tend to outperform their cross-prompt counterparts in terms of predictive accuracy; (2) prompt-specific models frequently exhibit a greater bias towards students of different economic statuses compared to cross-prompt models; (3) in the pursuit of generalizability, traditional machine learning models (e.g., SVM) coupled with carefully engineered features hold greater potential for achieving both high accuracy and fairness than complex neural network models.

Introduction

In education, writing is a prevalent pedagogical practice employed by teachers and instructors to enhance student learning (Defazio et al. 2010). Yet, the timely evaluation of students' essays or responses represents a formidable challenge, consuming considerable time and cognitive effort for educators. Recognizing the need to alleviate this burden, Automatic Essay Scoring (AES) has emerged, which refers to the process of using machine learning techniques to evaluate and assign scores to student-authored essays or responses (Chodorow and Burstein 2004). By automating this assessment process, educators can better focus on refining their teaching strategies, ultimately enabling a more efficient and effective learning experience for students.

Given the significant potential of AES, substantial efforts have been directed towards this field (Larkey 1998; Miltsakaki and Kukich 2004; Chen and He 2013; McNamara et al. 2015). It is important to highlight that a common objective shared among existing AES investigations is the pursuit of optimal predictive accuracy, i.e., correctly assessing and assigning scores to essays as many as possible. For instance, an early study (Zesch, Wojatzki, and Scholten-Akoun 2015) enhanced the training of an AES model based on Support Vector Machine (SVM) through a comprehensive feature set encompassing key linguistic attributes crucial for essay quality assessment (e.g., word n-gram features, cohesion features, and syntax features). The advancements in deep neural networks have spurred endeavors to further elevate predictive accuracy (Taghipour and Ng 2016; Alikaniotis, Yannakoudakis, and Rei 2016). These range from crafting dedicated scoring models based on different neural network architectures (e.g., Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM)) to harnessing pretrained large language models (e.g., BERT (Devlin et al. 2019)). Noteworthy is that the aforementioned accuracy-focused studies were frequently operated within the *prompt-specific* context, i.e., the AES models were developed and evaluated using labeled data exclusive to the intended target prompt. Nevertheless, obtaining such labeled data may not always be feasible, given its potential scarcity or the significant expenses and time required for its preparation. This has led to a recent trend in AES that centers on augmenting model generalizability in a *cross-prompt* setting, i.e., building AES models based on pre-existing data sourced from non-target prompt and subsequently assessing their applicability to the desired target prompt (Jin et al. 2018; Ridley et al. 2020; Li, Chen, and Nie 2020).

While significant progress has been made, current research falls short in offering comprehensive insights to real-world educators on effectively balancing various factors crucial for constructing effective AES models. For instance, though the generalizability of cross-prompt models is a de-

*Corresponding authors.

sirable trait to have, it is not the only trait that educators consider when determining which model they should use in practice. If the predictive accuracy of a cross-prompt model significantly lags behind that of a prompt-specific model, educators might favor the prompt-specific model, even when they acknowledge the associated costs of preparing tailored training data. However, the comparison between prompt-specific and cross-prompt AES models regarding their predictive accuracy remains largely unexplored in the existing literature. Beyond accuracy and generalizability, the predictive fairness of AES models has garnered increasing attention from educators. Here, predictive fairness entails ensuring that the essay score predictions generated by an AES model are impartial and unbiased across diverse student groups characterized by varying sensitive attributes such as gender and age. Undoubtedly, any bias hidden behind machine learning models can lead to unfair and discriminatory outcomes towards students, and thus should be addressed. Despite its recognized significance, the fairness of AES methods has received limited investigation within existing studies.

Hence, this study aimed to systematically investigate the intricate relationship between an AES model's accuracy and fairness, and generalizability, shedding light on practical insights for real-world educators to develop effective AES models to better support their teaching practices. Formally, this study was guided by the following two **R**esearch **Q**uestions:

RQ1 What is the performance difference between the prompt-specific and cross-prompt AES methods in terms of predictive *accuracy*?

RQ2 What is the performance difference between the prompt-specific and cross-prompt AES methods in terms of predictive *fairness*?

To answer the RQs, we chose a publicly available dataset consisting of over 25,000 argumentative essays with holistic essay scores from 15 distinct prompts. Notably, this dataset provides various demographic details pertaining to students, including gender, economic status, disability status, English language learner status, and race. This rich dataset facilitated our exploration of AES model biases through various demographic lenses. To ensure a comprehensive evaluation, we extensively reviewed the existing AES literature and selected nine prominent methods in the field, with five from the prompt-specific category and four from the cross-prompt category. Subsequently, we replicated all nine methods and evaluate their accuracy and fairness measured by seven different metrics. The details are provided in Section Methods. We have publicly released our code[1].

In summary, this study contributed to the AES literature with the following main findings and insights:

- Prompt-specific models tend to outperform their cross-prompt counterparts, with the performance gap ranging from 18.06% to 25.61% depending on the evaluation metrics used;
- In the cross-prompt setting, simple models (e.g., those based on well-investigated machine learning models like SVM) often excel in adequately identifying the characteristics of quality essays compared to complex models based on deep neural networks;
- Students' economic status emerges as the major attribute which frequently suffers from the predictive bias of existing AES models;
- Prompt-specific models frequently exhibit more bias towards students of different economic statuses compared to cross-prompt models;
- In the pursuit of generalizability, traditional machine learning models with carefully handcrafted features can achieve both high accuracy and fairness.

[1]https://github.com/CarsonYang518/AAAI24-AES-AFG

Related Work

Automatic Essay Scoring

The AES studies that are most relevant to our work can be broadly categorized into two groups, namely *prompt-specific* and *cross-prompt*, as briefly summarized below.

Prompt-specific AES. The initial explorations within this category predominantly relied on traditional machine learning techniques such as Bayesian Linear Regression, ν-SVM, and Random Forests (RF) (Rudner and Liang 2002; Cozma, Butnaru, and Ionescu 2018; Chen and He 2013). To equip an AES model with the ability to accurately evaluate the quality of an essay, these studies often placed significant emphasis on the manual crafting of meaningful textual features as input to train the model. For instance, (Zesch, Wojatzki, and Scholten-Akoun 2015) empowered the training of a SVM-based scoring model by engineering an extensive set of linguistic features, encompassing critical aspects such as essay length, syntax, and coherence. Inspired by the strides made in deep learning techniques to address diverse natural language processing challenges, several studies have been dedicated to applying these methodologies to tackle AES (Taghipour and Ng 2016; Dong and Zhang 2016; Dong, Zhang, and Yang 2017; Tay et al. 2018). In contrast to traditional machine learning models, deep learning models dispense with the need for hand-crafted features, proficiently extracting such features from raw textual data. The focus of these deep learning studies often centers on the use of diverse deep neural network architectures that are adept at capturing distinct textual attributes within essays to facilitate subsequent grading. For example, CNNs are harnessed to discern local textual dependencies, while LSTM networks are employed to capture sequential dependencies (Taghipour and Ng 2016). Hierarchical network structures are used to capture both word-level and sentence-level dependencies (Dong and Zhang 2016), and attention mechanisms are deployed to pinpoint pivotal words or sentences crucial for determining essay quality (Dong, Zhang, and Yang 2017). Besides, the recent advancements in pre-trained large language models (e.g., BERT (Devlin et al. 2019)) have spurred researchers to leverage these cutting-edge tools for automating essay assessment (Rodriguez, Jafari, and Ormerod 2019; Yang et al. 2020; Mizumoto and Eguchi 2023).

Cross-prompt AES. The common assumption held by this strand of studies is that, though the labeled data pertinent to the target prompt is unavailable to train a prompt-specific model, the quality of an essay can somewhat be revealed by features that are important across all prompts (e.g., the number of grammatical errors contained in the essay). Therefore, these studies often endeavored to craft such features to empower the training of an AES model (Zesch, Wojatzki, and Scholten-Akoun 2015; Ridley et al. 2020). For example, (Zesch, Wojatzki, and Scholten-Akoun 2015) engineered weakly prompt-dependent features from 13 categories including the number of grammar errors, type-token-ratio, and readability score to train a SVM-based scoring model, whose predictive accuracy was up to 0.6856 measured by the metric of Quadratic Weighted Kappa (QWK). Building on top of this idea, (Jin et al. 2018) further proposed that the model built using the weakly prompt-dependent features could be used to accurately assign scores to certain essays, i.e., those receiving extremely high or low scores, and these essays together with their predicted scores can be used to further train a prompt-specific model. Specifically, (Jin et al. 2018) first trained a RankSVM model (Joachims 2002) powered by weakly prompt-dependent features based on the labeled data collected from non-target prompts. Then, the RankSVM model was employed to identify a set of essays that were of extremely high or low scores from the target prompt, which were further used as input to train a prompt-specific model based on two-layer LSTM neural networks. Similar studies were presented in (Liu and Ding 2021; Li, Chen, and Nie 2020).

Despite substantial endeavors aimed at bolstering the generalizability of AES models, a comprehensive evaluation and comparison of the predictive accuracy disparity between cross-prompt models and prompt-specific counterparts remain absent in existing studies. This unavoidably hinders educators from understanding the inherent trade-off between accuracy and generalizability they might encounter when devising real-world AES models. More importantly, none of the aforementioned studies have undertaken an evaluation of the fairness aspect of existing AES models.

Fair Machine Learning in Education

Given the important role played by machine learning in supporting teaching and learning, an increasing amount of attention has been given to the predictive bias of machine learning techniques used in education. According to a recent survey (Li et al. 2023), there have been only 49 peer-reviewed empirical papers on this topic published after 2010. These papers mostly centered around the tasks such as predicting students' course performance or their likelihood of dropping out from a course. To our knowledge, there are only two papers that diagnosed the predictive bias displayed by AES models, even though the importance of this task has been pointed out as early as in 2012 (Williamson, Xi, and Breyer 2012). Specifically, (Litman et al. 2021) evaluated the fairness of three prompt-specific models, i.e., one based on the RF model with handcrafted features, one based on CNN-LSTM-Attention with textual features, and one based on CNN-LSTM-Attention with hybrid features. Nonetheless, this study is limited in that it did not include any cross-prompt models and the findings were derived based on a private dataset consisting of data from only one prompt, which inherently hinders their reproducibility and generalizability in similar scenarios. In contrast, we delivered a more comprehensive evaluation by including nine prominent methods that encompass both the prompt-specific and cross-prompt settings, and the evaluation was based on a larger-scale public dataset collected from 15 distinct prompts. Additionally, (Doewes et al. 2022) measured individual fairness in AES while our work focused on group fairness.

Methods

Datasets

A major obstacle hindering the exploration of fair AES is the absence of demographic information within widely used datasets for AES research. To our knowledge, only two public datasets contain students' demographic information: the ELLIPSE Corpus and the PERSUADE 2.0 corpus [2]. For our evaluation, we chose the PERSUADE 2.0 corpus due to its larger dataset size, approximately five times that of the ELLIPSE Corpus. This will adequately fulfill the requirements of training data quantity for AES models based deep learning techniques, as described in Section Models.

The PERSUADE 2.0 corpus originally consists of over 25,000 argumentative essays written by students from 6th to 12th grade in the US for 15 different prompts (Crossley et al. 2022). The holistic essay scores, which serve as the ground truth for this study's predictions, were assigned by human raters who underwent training on a scoring rubric employed in the standardized Scholastic Aptitude Test (SAT) in the US. These holistic scores span from 1 to 6, denoting low to high quality, with increments of 1. Importantly, the dataset encompasses five demographic attributes of the students, including *gender* (male vs. female), *race/ethnicity* (e.g., White, Asian, Black), *economic status* (economically disadvantaged vs. non-economically disadvantaged), *English language learner status* (native English speakers vs. non-native English speakers), and *disability status* (students with disabilities and those without). All these demographic attributes were considered to address RQ2. The details of the dataset are described in (Crossley et al. 2023).

Models

Following prior research (Tay et al. 2018; Ridley et al. 2020; Jin et al. 2018; Cozma, Butnaru, and Ionescu 2018; Yang et al. 2020), we treated the prediction of an essay's score as a regression problem. To ensure a comprehensive evaluation, we conducted an extensive review on the existing AES literature, after which we chose five representative ***prompt-specific*** methods, as described below:

- **`SVM (Full)`** (Zesch, Wojatzki, and Scholten-Akoun 2015), which aims to adequately empower the training of a SVM-based scoring model by carefully engineering a comprehensive set of features from raw essay text. The authors distinguished two types of features, namely (i)

[2] https://github.com/scrosseye/persuade_corpus_2.0

strongly prompt-dependent ones, i.e., those highly associated with a specific prompt such as word n-grams and essay length; and (ii) *weakly prompt-dependent* ones, i.e., those matter to the essay assessment of all prompts such as grammatical errors and readability.

- **SKIPFLOW-LSTM** (Tay et al. 2018), which is a pioneering attempt to incorporate features related to the coherence of an essay (i.e., the semantic similarity between different sentences) to train an AES based on an end-to-end neural network architecture. This architecture encompasses a neural tensor layer to capture the relationship between two LSTM outputs, with the goal of automatically extracting coherence features for essay scoring.
- **CNN-LSTM-ATT** (Dong, Zhang, and Yang 2017), which is the first study to employ a neural hierarchical sentence-document architecture for AES. Specifically, this study used CNN to capture the word relations in a sentence and then LSTM to capture the sentence relations in an essay. Besides, the attention mechanism was applied to identify crucial words and sentences for assessing essay quality.
- **R^{2}BERT** (Yang et al. 2020), which is a pioneering attempt to combine the methodologies of regression and ranking in AES. Specifically, a hybrid loss with the dynamic weights of mean square error loss (i.e., regression loss) and batch-wise ListNet loss (i.e., ranking loss) is applied to fine-tune the BERT for AES.
- **BERT (3 Layers)** (Rodriguez, Jafari, and Ormerod 2019), which aims to mitigate overfitting by only using the initial three layers of BERT to produce essay representations for subsequent scoring. This configuration setting was demonstrated to yield the optimal performance after extensive experimentation of various alternative configurations and training techniques for BERT.

Similarly, we chose four representative ***cross-prompt*** methods, as described below:

- **SVM (Reduced)** (Zesch, Wojatzki, and Scholten-Akoun 2015), which is similar to SVM (Full) described above, but only using the weakly prompt-dependent features for model training.
- **RankSVM** (Chen, Xu, and He 2014), which is a representative ranking-based method for AES. A RankSVM is first trained using pair-wise essays ordered by ground-truth scores. Then, the constructed RankSVM is used to generate intermediate scores for ranking the essays, and such intermediate scores are subsequently mapped to a pre-defined scoring scale to generate the essay scores.
- **PAES** (Ridley et al. 2020), which is similar to CNN-LSTM-ATT mentioned above. However, Part-of-Speech (POS) embeddings are used here rather than word embeddings, because POS embeddings are assumed to be more effective in generating a generalized representation of an essay. Besides, this method incorporates certain weakly task-dependent features to train the AES model.
- **TDNN** (Jin et al. 2018), which introduces a pioneering two-step approach for cross-prompt essay scoring. Firstly, a RankSVM model is constructed as described above. Secondly, the RankSVM model is used to assign scores to essays in the desired target prompt, among which the essays receiving extremely high or low scores are further used to train a LSTM neural network for prompt-specific essay scoring.

Experimental Setup

Data Preprocessing. The essays without corresponding student demographic information were removed, which resulted in a total of 20,626 essays spanning 12 distinct prompts for our evaluation. Notably, four out of the five demographic attributes are in a binary form (e.g., male vs. female), except for *race/ethnicity*, which contains six values including White, Hispanic, Black, Asian, American Indian, and Other. As guided by (Hutt et al. 2019; Bayer, Hlosta, and Fernandez 2021), White students are regarded as the privileged group, we converted this attribute to binary values of White vs. Non-White. In line with existing studies in the field (Litman et al. 2021), we treated students who are either male, White, economically advantaged, native English speakers, or without disabilities as the privileged group and the others as the non-privileged group to measure the fairness of AES models.

Feature Engineering. The handcrafted features of the models SVM (Full/Reduced), RankSVM, PAES, and TDNN (as specified in Section Models were derived using NLTK (Loper and Bird 2002), Stanza (Qi et al. 2020), and spaCy (Honnibal and Montani 2017). In line with previous studies (Jin et al. 2018; Ridley et al. 2020), we standardized all handcrafted features, adjusting their means to 0 and standard deviations to 1.0. The details about handcrafted features can be found in the Appendix.

Model Construction. We employed both TensorFlow (Paszke et al. 2019) and PyTorch (Abadi et al. 2015) frameworks for implementing the deep learning models, namely SKIPFLOW-LSTM, CNN-LSTM-ATT, R^{2}BERT, BERT (3 Layers), PAES, and TDNN (as detailed in Section Models). For traditional machine learning models (e.g., SVM), we employed Scikit-learn (Pedregosa et al. 2011). For RankSVM, we employed the SVMs library [3]. All the model hyperparameters were set following the guidelines specified in the original papers and can be found in the Appendix. All the model training and evaluation were performed on Google Colab Pro with 16 GB of RAM and an NVIDIA Tesla T4 GPU.

Evaluation Procedure. In previous studies (Mathias and Bhattacharyya 2018; Cozma, Butnaru, and Ionescu 2018; Dong, Zhang, and Yang 2017; McNamara et al. 2015), the evaluation of prompt-specific methods often involved 5-fold cross-validation. As for the evaluation of cross-prompt methods, a prompt-wise cross-validation approach is commonly employed (Ridley et al. 2020; Jin et al. 2018; Liu and Ding 2021), where essays corresponding to a target prompt are held out for testing, while the remaining essays of other prompt are utilized as training data. We adopt the same evaluation procedure as previous studies. By doing this, all the

[3] https://www.cs.cornell.edu/people/tj/svm_light/svm_rank.html

Types	Metrics	↑QWK	↓MAE	↑PCC	↑QWK	↓MAE	↑PCC	↑QWK	↓MAE	↑PCC	↑QWK	↓MAE	↑PCC
	Methods	**Prompt 1**			**Prompt 2**			**Prompt 3**			**Prompt 4**		
PS	SVM (Full)	0.823	0.456	0.845	0.770	0.495	0.819	0.772	0.528	0.816	0.717	0.473	0.758
	SKIPFLOW-LSTM	0.781	0.527	0.798	0.838	0.465	0.849	0.800	0.532	0.813	0.713	0.489	0.747
	CNN-LSTM-ATT	0.838	0.473	0.846	0.855	0.451	0.864	0.824	0.515	0.832	0.777	**0.444**	**0.796**
	R^2BERT	0.822	0.530	0.826	0.849	**0.422**	0.857	0.831	**0.469**	0.837	0.740	0.464	0.751
	BERT (3 Layers)	**0.858**	**0.452**	**0.861**	**0.866**	0.435	**0.869**	**0.844**	0.501	**0.850**	**0.787**	0.451	0.792
CP	SVM (Reduced)	0.835	0.462	0.851	0.774	0.520	0.806	0.761	0.566	0.785	0.759	0.475	0.773
	Rank-SVM	0.747	0.573	0.798	0.587	0.710	0.702	0.499	0.820	0.678	0.584	0.602	0.667
	PAES	0.809	0.567	0.833	0.760	0.566	0.783	0.730	0.669	0.754	0.666	0.719	0.719
	TDNN	0.732	0.692	0.789	0.609	0.690	0.646	0.560	0.721	0.642	0.536	0.602	0.599
		Prompt 5			**Prompt 6**			**Prompt 7**			**Prompt 8**		
PS	SVM (Full)	0.753	0.475	0.790	0.783	0.449	0.813	0.732	0.425	0.774	0.704	0.454	0.747
	SKIPFLOW-LSTM	0.674	0.552	0.708	0.773	0.487	0.793	0.730	0.451	0.752	0.728	0.451	0.745
	CNN-LSTM-ATT	0.775	0.477	**0.793**	0.813	0.449	0.826	0.776	0.414	0.795	0.763	0.434	0.779
	R^2BERT	0.708	0.556	0.727	0.802	0.473	0.815	0.783	0.389	0.788	0.768	0.433	0.772
	BERT (3 Layers)	**0.783**	0.486	0.792	**0.840**	**0.428**	**0.845**	**0.806**	**0.393**	**0.811**	**0.788**	**0.417**	**0.793**
CP	SVM (Reduced)	0.764	**0.474**	0.789	0.822	0.437	0.833	0.747	0.438	0.790	0.740	0.436	0.771
	Rank-SVM	0.691	0.551	0.718	0.723	0.517	0.743	0.521	0.537	0.546	0.522	0.581	0.563
	PAES	0.690	0.661	0.751	0.749	0.606	0.809	0.711	0.548	0.747	0.662	0.620	0.742
	TDNN	0.620	0.608	0.641	0.727	0.581	0.733	0.090	0.849	0.157	0.225	0.743	0.317

Table 1: The predictive accuracy of the selected AES methods in each prompt. PS represents Prompt-Specific. CP represents Cross-Prompt. Bold values represent the best performance in a metric. The signs ↑ and ↓ indicate whether a higher (↑) or lower (↓) value is more preferred in a metric.

essays contained in a target prompt were scored by prompt-specific and cross-prompt methods, which enabled us to directly compare their performance.

Evaluation Metric

To measure *accuracy*, we adopt three commonly used metrics in existing AES literature (Lagakis and Demetriadis 2021): Quadratic Weighted Kappa (QWK), Mean Absolute Error (MAE), and Pearson Correlation Coefficient (PCC). Although QWK is designed for categorical variables, we adapted it for our regression task by utilizing a modified version suited for continuous values (Haberman 2019).

To measure *fairness*, we aligned with previous studies (Loukina, Madnani, and Zechner 2019; Litman et al. 2021) and adopt three metrics to measure to what extent the predictive errors of an AES model towards different student groups can be attributed to their demographic traits:

- Overall Score Accuracy (OSA), which measures the parity of an AES model in terms of the variance between its predicted scores and the ground-truth scores that can be explained by students' demographic attributes. Specifically, OSA represents the scores given by an AES model and the human rater with S and H, respectively. Then, a linear regression is constructed with $(S-H)^2$ as the dependent variable and demographic attributes as the independent variable. OSA is calculated as the R^2 of this regression model.
- Overall Score Difference (OSD), which is similar to OSA, but with $S-H$ (instead of $(S-H)^2$) to construct the regression model. This is designed to capture any "overestimation" or "underestimation" displayed by an AES model towards any group of students (e.g., whether the AES model tends to assign higher scores to essays written by male students while their female counterparts often receive lower scores).
- Conditional Score Difference (CSD), which is similar to OSD, takes a step further by accounting for students' language proficiency, approximated by their ground-truth essay scores. This is achieved by constructing two regression models with $S-H$ as the dependent variable, first with H as the independent variable, and then with both H and demographic attributes. CSD is calculated as the difference between R^2 of these two regression models.

The larger the OSA/OSD/CSD, the more bias an AES model has. We employed ANOVA to assess whether the results of CSD were statistically significant. In addition to using OSA, OSD, and CSD to explain the scoring error variance across different demographic groups, we measured fairness from the scale perspective by adopting Mean Absolute Error Difference (MAED) (Sun, Fung, and Haghighat 2022), which calculates the difference between the MAE of the privileged and unprivileged groups. Positive MAED values indicate that the AES model holds bias towards the privileged group while negative values indicate bias towards the non-privileged group. That is, the closer a MAED is to 0, the more fair an AES model is. All the evaluation metrics were calculated using RSMTool (Madnani and Loukina 2016).

Results

Results on RQ1

Limited space allows us to display tables for some prompts and one demographic; access the full tables here [4].

The predictive accuracy of the nine selected AES methods in each prompt is detailed in Table 1, which is further

[4]https://bit.ly/AAAI24-AES-AFG

Types	Metrics	OSA	OSD	CSD	MAED	OSA	OSD	CSD	MAED	OSA	OSD	CSD	MAED
	Methods	**Prompt 1**				**Prompt 2**				**Prompt 3**			
PS	SVM (Full)	ns	ns	ns	0.052	ns	0.022	ns	-0.087	ns	0.045	ns	-0.011
	SKIPFLOW-LSTM	0.111	0.112	0.014	-0.512	0.078	0.079	0.005	-0.459	0.051	0.052	ns	-0.386
	CNN-LSTM-ATT	ns	ns	0.020	0.045	ns	ns	0.016	-0.043	ns	ns	ns	0.033
	R^2BERT	ns	ns	0.022	0.010	ns	ns	0.023	-0.084	ns	ns	ns	0.001
	BERT (3 Layers)	ns	ns	0.015	0.037	ns	ns	ns	-0.029	ns	ns	ns	0.015
CP	SVM (Reduced)	ns	ns	ns	0.015	ns	ns	ns	0.022	ns	0.031	ns	0.035
	Rank-SVM	ns	ns	ns	0.016	ns	0.025	ns	0.107	0.015	0.05	ns	0.180
	PAES	ns	ns	ns	0.087	ns	ns	ns	0.042	ns	ns	ns	0.048
	TDNN	ns	ns	ns	0.099	ns	0.022	ns	0.057	ns	0.054	ns	0.036
		Prompt 4				**Prompt 5**				**Prompt 6**			
PS	SVM (Full)	ns	ns	ns	-0.008	ns	ns	ns	0.031	ns	ns	ns	0.040
	SKIPFLOW-LSTM	0.043	0.043	ns	-0.243	0.051	0.051	ns	-0.293	0.104	0.111	ns	-0.494
	CNN-LSTM-ATT	ns	ns	ns	0.012	ns	ns	ns	0.022	ns	ns	0.012	0.008
	R^2BERT	ns	ns	ns	-0.031	ns	ns	ns	-0.001	ns	ns	0.022	-0.035
	BERT (3 Layers)	ns	ns	ns	-0.005	ns	ns	ns	0.038	ns	ns	0.026	-0.023
CP	SVM (Reduced)	ns	ns	ns	0.001	ns	ns	ns	0.026	ns	ns	ns	0.016
	Rank-SVM	ns	ns	ns	0.078	ns	ns	ns	0.001	ns	ns	ns	0.090
	PAES	ns	ns	0.028	-0.012	ns	ns	ns	0.072	ns	ns	ns	0.084
	TDNN	ns	ns	0.005	0.033	ns	ns	ns	0.041	ns	ns	ns	0.012
		Prompt 7				**Prompt 8**				**Prompt 9**			
PS	SVM (Full)	ns	ns	ns	-0.037	ns	ns	ns	-0.014	ns	ns	ns	-0.023
	SKIPFLOW-LSTM	ns	ns	ns	-0.239	0.038	0.04	ns	-0.245	0.036	0.038	ns	-0.259
	CNN-LSTM-ATT	ns	ns	ns	-0.008	ns	ns	ns	-0.003	ns	ns	ns	-0.001
	R^2BERT	ns	ns	ns	-0.014	ns	ns	ns	-0.002	ns	ns	ns	-0.030
	BERT (3 Layers)	ns	ns	ns	-0.008	ns	ns	0.015	-0.018	ns	ns	ns	-0.043
CP	SVM (Reduced)	ns	ns	ns	-0.038	ns	ns	ns	-0.006	ns	ns	ns	-0.059
	Rank-SVM	ns	ns	ns	0.017	ns	0.015	ns	0.066	ns	ns	ns	-0.002
	PAES	ns	ns	0.022	-0.005	ns	ns	0.009	-0.036	ns	ns	ns	-0.081
	TDNN	ns	0.048	ns	-0.221	ns	0.03	ns	-0.149	ns	ns	ns	-0.133

Table 2: The predictive fairness of the selected AES methods for Economic Status. S represents Prompt-Specific. CP represents Cross-Prompt. The 'ns' label indicates non-significant results ($p < 0.05$). Lower values indicate a higher level of fairness.

averaged and presented in Table 3. Based on these tables, two interesting observations can be made.

Firstly, prompt-specific models generally outperform cross-prompt models. As shown in Table 3, on average, QWK exhibits a 25.61% increase, the MAE shows a reduction of 23.43%, and the PCC demonstrates an enhancement of 18.06%. When comparing the best-performing prompt-specific model `BERT (3 Layers)`) and its best-performing cross-prompt counterpart (i.e., `SVM (Reduced)`), the performance gap is 9.00% in QWK, 8.71% in MAE, and 5.42% in PCC. On the other hand, in line with previous research (Zesch, Wojatzki, and Scholten-Akoun 2015; Cozma, Butnaru, and Ionescu 2018), we observed that prompt-specific models tended to display greater robustness compared to cross-prompt ones, as evidenced by the variances shown in Table 3. This is due to the more challenging nature of the cross-prompt essay scoring as it can not leverage prompt-specific features (e.g., n-grams) that directly contribute to the accurate evaluation of an essay.

Secondly, when scrutinizing the prompt-specific models, we observe that models based on deep neural networks are consistently superior to those based on traditional machine learning techniques. For instance, the best performing model `BERT (3 Layers)` achieved an average performance of up to 0.811 (QWK), 0.440 (MAE), and 0.817 (PCC). Notably, this model also achieved the highest level of robustness as indicated by the lowest variances (as low as 0.001) among all the prompt-specific models. However, when scrutinizing the cross-prompt models, we have the contrary finding, i.e., the traditional machine learning method `SVM (Reduced)` exhibits the highest performance compared to all the other deep learning methods (namely PAES and TDNN). This implies that, in the cross-prompt setting, simple models can effectively discern significant patterns of quality essays by using weakly prompt-dependent features, while complex models based on deep neural networks (e.g., TDNN, which is an advanced version of RankSVM) have the tendency to overfit non-target-prompt essays, thereby diminishing their ability to generalize effectively.

Results on RQ2

The predictive fairness was evaluated by using all the five available demographic attributes, among which we observed that an AES model's bias is frequently associated with a student's *economic status*. The predictive bias of AES methods in different prompts is given in Table 2, which are further

Types	Methods	↑QWK	σ^2	↓MAE	σ^2	↑PCC	σ^2	OSA	OSD	CSD	MAED
PS	SVM(Full)	0.733	0.004	0.466	0.001	0.781	0.002	0	2	0	-0.006
	SKIPFLOW-LSTM	0.739	0.003	0.497	0.001	0.763	0.002	10	10	2	-0.312
	CNN-LSTM-ATT	0.789	0.002	0.462	0.002	0.805	0.001	0	0	4	0.004
	R^2BERT	0.777	0.002	0.456	0.003	0.786	0.002	0	0	3	-0.027
	BERT(3 Layers)	**0.811**	0.001	**0.440**	0.001	**0.817**	0.001	0	0	3	-0.005
CP	SVM(Reduced)	0.744	0.006	0.482	0.004	0.775	0.004	0	1	0	0.001
	Rank-SVM	0.579	0.016	0.613	0.009	0.643	0.014	2	4	0	0.063
	PAES	0.680	0.007	0.643	0.011	0.734	0.005	0	0	3	0.020
	TDNN	0.450	0.060	0.687	0.006	0.528	0.041	0	4	2	-0.009

Table 3: The average accuracy performance and overall fairness performance for Economic Status of the selected AES methods across all prompts. σ^2 represents variance. PS represents Prompt-Specific. CP represents Cross-Prompt. Bold values represent the best performance in a metric. The signs ↑ and ↓ indicate whether a higher (↑) or lower (↓) value is more preferred in a metric. Cells in OSA, OSD, and CSD denote the number of prompts in which an AES method was diagnosed to have predictive bias, e.g., the number of cells with values other than 'ns' in Table 2. Cells in MAED represent the average MAED of all prompts.

summarized and presented in Table 3. This aligns with the findings presented in previous studies (Abdu-Raheem 2015), i.e., there exists a relationship between students' academic achievements and their parents' socio-economic status. On the other hand, *gender* is the attribute in which AES models display relatively fewer biases in our case.

When delving into the results of economic status presented in Table 2 and Table 3, we observe that prompt-specific models generally displayed more bias compared to their cross-prompt counterparts. Specifically, when centering on the metrics of OSA, OSD, and CSD, the average number of prompts that the prompt-specific models were diagnosed to have bias is greater than that of the cross-prompt models, namely 2.0 vs. 0.5 in OSA, 2.4 vs. 2.25 in OSD, and 2.4 vs. 1.25 in CSD. On the other hand, when calculating the average of the absolute MAED values, the performance of prompt-specific models is 2.63% higher than that of the cross-prompt models. It should be noted that cross-prompt models tended to favor the non-privileged group (i.e., three out of four models displayed positive MAED values) while prompt-specific models were more likely to favor the privileged group (i.e., four out of the five models displayed negative MAED values).

When scrutinizing the fairness displayed by individual models in the prompt-specific setting, we observe that `SVM (Full)` is superior to the other methods, with only two prompts detected with bias measured in OSD and a minimal MAED value of -0.006. This model is followed by `BERT (3 Layers)`, which was diagnosed to be biased in three prompts and with a MAED value of -0.005. Recall the RQ1 results presented in Table 1 and Table 3, `BERT (3 Layers)` demonstrated the highest predictive accuracy and robustness. This further strengthens the superiority of AES models based on meticulously fine-tuned pre-trained large language models in the prompt-specific setting, which can simultaneously achieve high accuracy and fairness. A similar conclusion can be drawn for the cross-prompt models. That is when pursuing generalizability, simple models based on well-investigated machine learning models such as SVM coupled with informative hand-crafted features might be preferable to complex models based on deep neural networks to achieve not only accurate but also fair evaluation.

Discussion and Conclusion

To better support instructors and educators in selecting approximate AES models, we carefully selected nine representative AES approaches, covering both prompt-specific and cross-prompt categories. Subsequently, we evaluated the effectiveness of these methods on an open-sourced dataset with five demographic attributes using seven distinct metrics that account for both accuracy and fairness. Upon scrutinizing the results, we derive the subsequent implications and acknowledge the limitations of our study.

Implications. Firstly, the results reveal a 9.00% QWK gap, 8.71% MAE gap, and 5.42% PCC gap between the top-performing prompt-specific model (`BERT (3 Layers)`) and the best cross-prompt model (`SVM (Reduced)`). This suggests that choosing `SVM (Reduced)` could improve generalizability, although with some accuracy trade-offs. Secondly, `BERT (3 Layers)` excels in fairness (MAED of -0.005, just 0.001 apart from the best) and achieves the highest accuracy in prompt-specific settings, making it a strong recommendation for such settings. `CNN-LSTM-ATT` delivers top fairness (MAED of 0.004) and the second-best accuracy (2.7% QWK decrease from the best) in prompt-specific settings, making it another strong recommendation.

Limitations. We acknowledged the following limitations of our study. Firstly, our experiments were restricted to a single dataset, underscoring the need to enhance the broader applicability of our findings through the inclusion of supplementary datasets in our evaluation process. Secondly, our analysis was predominantly centered around evaluating fairness, without providing definite solutions for addressing the identified fairness disparities. In future research, our emphasis will be on mitigating model unfairness while upholding an acceptable level of accuracy.

Acknowledgments

This work is supported in part by National Key R&D Program of China (2022YFC3303603), Australian Research Council (DP220101209), NSFC (62077028, 62377028), Key Laboratory of Smart Education of Guangdong Higher Education Institutes, Jinan University (2022LSYS003).

References

Abadi, M.; Agarwal, A.; Barham, P.; Brevdo, E.; Chen, Z.; Citro, C.; Corrado, G. S.; Davis, A.; Dean, J.; Devin, M.; Ghemawat, S.; Goodfellow, I.; Harp, A.; Irving, G.; Isard, M.; Jia, Y.; Jozefowicz, R.; Kaiser, L.; Kudlur, M.; Levenberg, J.; Mané, D.; Monga, R.; Moore, S.; Murray, D.; Olah, C.; Schuster, M.; Shlens, J.; Steiner, B.; Sutskever, I.; Talwar, K.; Tucker, P.; Vanhoucke, V.; Vasudevan, V.; Viégas, F.; Vinyals, O.; Warden, P.; Wattenberg, M.; Wicke, M.; Yu, Y.; and Zheng, X. 2015. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. Software available from tensorflow.org.

Abdu-Raheem, B. 2015. Parents' Socio-Economic Status as Predictor of Secondary School Students' Academic Performance in Ekiti State, Nigeria. *Journal of Education and practice*, 6(1): 123–128.

Alikaniotis, D.; Yannakoudakis, H.; and Rei, M. 2016. Automatic Text Scoring Using Neural Networks. *CoRR*, abs/1606.04289.

Bayer, V.; Hlosta, M.; and Fernandez, M. 2021. Learning analytics and fairness: do existing algorithms serve everyone equally? In *International Conference on Artificial Intelligence in Education*, 71–75. Springer.

Chen, H.; and He, B. 2013. Automated Essay Scoring by Maximizing Human-Machine Agreement. In *Conference on Empirical Methods in Natural Language Processing*.

Chen, H.; Xu, J.; and He, B. 2014. Automated essay scoring by capturing relative writing quality. *The Computer Journal*, 57(9): 1318–1330.

Chodorow, M.; and Burstein, J. 2004. Beyond essay length: evaluating e-rater®'s performance on toefl® essays. *ETS Research Report Series*, 2004(1): i–38.

Cozma, M.; Butnaru, A. M.; and Ionescu, R. T. 2018. Automated essay scoring with string kernels and word embeddings. *arXiv preprint arXiv:1804.07954*.

Crossley, S.; Baffour, P.; Yu, T.; Franklin, A.; Benner, M.; and Boser, U. 2023. A large-scale corpus for assessing written argumentation: PERSUADE 2.0.

Crossley, S. A.; Baffour, P.; Tian, Y.; Picou, A.; Benner, M.; and Boser, U. 2022. The persuasive essays for rating, selecting, and understanding argumentative and discourse elements (PERSUADE) corpus 1.0. *Assessing Writing*, 54: 100667.

Defazio, J.; Jones, J.; Tennant, F.; and Hook, S. A. 2010. Academic Literacy: The Importance and Impact of Writing across the Curriculum–A Case Study. *Journal of the Scholarship of Teaching and Learning*, 10(2): 34–47.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805.

Doewes, A.; Saxena, A.; Pei, Y.; and Pechenizkiy, M. 2022. Individual Fairness Evaluation for Automated Essay Scoring System. *International Educational Data Mining Society*.

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*, 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)*, 153–162.

Haberman, S. J. 2019. Measures of agreement versus measures of prediction accuracy. *ETS Research Report Series*, 2019(1): 1–23.

Honnibal, M.; and Montani, I. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. *To appear*, 7(1): 411–420.

Hutt, S.; Gardner, M.; Duckworth, A. L.; and D'Mello, S. K. 2019. Evaluating Fairness and Generalizability in Models Predicting On-Time Graduation from College Applications. *International Educational Data Mining Society*.

Jin, C.; He, B.; Hui, K.; and Sun, L. 2018. TDNN: a two-stage deep neural network for prompt-independent automated essay scoring. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1088–1097.

Joachims, T. 2002. Optimizing search engines using clickthrough data. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, 133–142.

Lagakis, P.; and Demetriadis, S. 2021. Automated essay scoring: A review of the field. In *2021 International Conference on Computer, Information and Telecommunication Systems (CITS)*, 1–6. IEEE.

Larkey, L. S. 1998. Automatic Essay Grading Using Text Categorization Techniques. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '98, 90–95. New York, NY, USA: Association for Computing Machinery. ISBN 1581130155.

Li, L.; Sha, L.; Li, Y.; Raković, M.; Rong, J.; Joksimovic, S.; Selwyn, N.; Gašević, D.; and Chen, G. 2023. Moral Machines or Tyranny of the Majority? A Systematic Review on Predictive Bias in Education. In *LAK23: 13th International Learning Analytics and Knowledge Conference*, 499–508.

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.

Litman, D.; Zhang, H.; Correnti, R.; Matsumura, L. C.; and Wang, E. 2021. A fairness evaluation of automated methods for scoring text evidence usage in writing. In *International Conference on Artificial Intelligence in Education*, 255–267. Springer.

Liu, C.; and Ding, G. 2021. MFDNN: Mixed Features Deep Neural Network Model for Prompt-independent Automated Essay Scoring. In *Proceedings of the 2021 4th International Conference on Algorithms, Computing and Artificial Intelligence*, 1–7.

Loper, E.; and Bird, S. 2002. Nltk: The natural language toolkit. *arXiv preprint cs/0205028*.

Loukina, A.; Madnani, N.; and Zechner, K. 2019. The many dimensions of algorithmic fairness in educational applications. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, 1–10.

Madnani, N.; and Loukina, A. 2016. RSMTool: A Collection of Tools for Building and Evaluating Automated Scoring Models. *Journal of Open Source Software*, 1(3).

Mathias, S.; and Bhattacharyya, P. 2018. ASAP++: Enriching the ASAP automated essay grading dataset with essay attribute scores. In *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)*.

McNamara, D. S.; Crossley, S. A.; Roscoe, R. D.; Allen, L. K.; and Dai, J. 2015. A hierarchical classification approach to automated essay scoring. *Assessing Writing*, 23: 35–59.

Miltsakaki, E.; and Kukich, K. 2004. Evaluation of Text Coherence for Electronic Essay Scoring Systems. *Nat. Lang. Eng.*, 10(1): 25–55.

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.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; and Duchesnay, E. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12: 2825–2830.

Qi, P.; Zhang, Y.; Zhang, Y.; Bolton, J.; and Manning, C. D. 2020. Stanza: A Python natural language processing toolkit for many human languages. *arXiv preprint arXiv:2003.07082*.

Ridley, R.; He, L.; Dai, X.; Huang, S.; and Chen, J. 2020. Prompt agnostic essay scorer: a domain generalization approach to cross-prompt automated essay scoring. *arXiv preprint arXiv:2008.01441*.

Rodriguez, P. U.; Jafari, A.; and Ormerod, C. M. 2019. Language models and automated essay scoring. *arXiv preprint arXiv:1909.09482*.

Rudner, L. M.; and Liang, T. 2002. Automated essay scoring using Bayes' theorem. *The Journal of Technology, Learning and Assessment*, 1(2).

Sun, Y.; Fung, B. C.; and Haghighat, F. 2022. In-Processing fairness improvement methods for regression Data-Driven building Models: Achieving uniform energy prediction. *Energy and Buildings*, 277: 112565.

Taghipour, K.; and Ng, H. T. 2016. A Neural Approach to Automated Essay Scoring. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 1882–1891. Austin, Texas: Association for Computational Linguistics.

Tay, Y.; Phan, M. C.; Tuan, L. A.; and Hui, S. C. 2018. SKIPFLOW: Incorporating Neural Coherence Features for End-to-End Automatic Text Scoring. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'18/IAAI'18/EAAI'18. AAAI Press. ISBN 978-1-57735-800-8.

Williamson, D. M.; Xi, X.; and Breyer, F. J. 2012. A framework for evaluation and use of automated scoring. *Educational measurement: issues and practice*, 31(1): 2–13.

Yang, R.; Cao, J.; Wen, Z.; Wu, Y.; and He, X. 2020. Enhancing automated essay scoring performance via fine-tuning pre-trained language models with combination of regression and ranking. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, 1560–1569.

Zesch, T.; Wojatzki, M.; and Scholten-Akoun, D. 2015. Task-independent features for automated essay grading. In *Proceedings of the tenth workshop on innovative use of NLP for building educational applications*, 224–232.