Can Language Models Guess Your Identity? Analyzing Demographic Biases in AI Essay Scoring

Alexander Kwako and **Christopher Ormerod**
Cambium Assessment Inc.
alexander.kwako@cambiumassessment.com
christopher.ormerod@cambiumassessment.com

Abstract

Large language models (LLMs) are increasingly used for automated scoring of student essays. However, these models may perpetuate societal biases if not carefully monitored. This study analyzes potential biases in an LLM (XLNet) trained to score persuasive student essays, based on data from the PERSUADE corpus. XLNet achieved strong performance based on quadratic weighted kappa, standardized mean difference, and exact agreement with human scores. Using available metadata, we performed analyses of scoring differences across gender, race/ethnicity, English language learning status, socioeconomic status, and disability status. Automated scores exhibited small magnifications of marginal differences in human scoring, favoring female students over males and White students over Black students. To further probe potential biases, we found that separate XLNet classifiers and XLNet hidden states weakly predicted demographic membership. Overall, results reinforce the need for continued fairness analyses as use of LLMs expands in education.

1 Introduction

As Large Language Models (LLM)s are increasingly used for Automated Essay Scoring (AES), it is crucial that we thoroughly analyze these systems for biases (Rodriguez et al., 2019). Given that LLMs are pretrained on large corpora, they have the potential to inherit biases embedded in the functions that predict word probabilities (Bhardwaj et al., 2021). If the potential biases are not monitored carefully with fairness in mind, they risk perpetuating and amplifying existing societal biases against vulnerable populations. Rigorous demographic analysis of AES systems help ensure they live up to principles of equity and fairness.

The Persuasive Essays for Rating, Selecting, and Understanding Argumentative and Discourse Elements (PERSUADE) corpus provides a valuable resource for analyzing bias in AES system (Crossley et al., 2022). PERSUADE contains over twenty-five thousand persuasive student essays that were annotated for argumentative elements in addition to holistic grades assigned by human raters. What makes this corpus ideal for the analysis of bias is the rich metadata about the students including gender, race, and other demographic indicators. This allows for in-depth analysis of an automated scoring system's performance on essays written by students of diverse demographic affiliations.

Our goal is to investigate potential biases in LLMs trained using conventional techniques that aim to replicate human-assigned holistic scores. After training the LLM scoring model, we evaluate whether or not automated scores introduce (or exacerbate) biases relative to human-assigned scores (Ormerod et al., 2022). After evaluating bias, we determine whether the set of features that the LLM uses for scoring also contains information relevant to demographic membership. In other words, can the LLM guess certain demographic characteristics based on the scoring model? Linear modeling using these features was recently used as evidence of model validity in AES (Ormerod, 2022). The novelty of this study lies in applying these techniques and showing their relevance to the analyses of bias in LLMs.

Broadly, our research aims are as follows:

1. Fine-tune an LLM to score students' essays and assess model performance.

2. Evaluate the fine-tuned LLM for biases relative to human-assigned scores, based on students' demographic affiliations.

3. Determine whether demographic affiliation can be predicted by the (hidden layer of the) fine-tuned LLM. As a helpful reference point, assess whether separate LLMs can be fine-tuned to predict demographic affiliation.

Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications, pages 78–86

These aims help us determine if LLMs are able to score students' essays fairly, and if demographic affiliations are an implicit feature of the scoring model.

2 Methods

2.1 Data

The PERSUADE dataset consists of a 25,488 essay responses to 15 prompts written by students from Grades 6 to 12.[1] Each essay was assigned a holistic essay score according to a rubric available with the dataset.

The prompts were administered to students within specific grades or grade-bands. For comparability with other studies, we used the same train-test split as was used in the original Kaggle competition; we created a development dataset (or *dev set*) from a random subset of the training data, for use in model selection, early stopping, and hyperparameter optimization. Table 1 shows sample sizes of train-dev-test splits, along with the average word count, for each prompt.

Demographic data was included for all prompts, but not all prompts included every demographic characteristic. For most prompts, however, we analyzed potential biases of the following demographic affiliations:

- Gender: M = Male and F = Female
- Race/ethnicity: W = White, L = Hispanic/Latino, B = Black/African American, A = Asian/Pacific Islander
- English Language Learners: ELL = Identified as an English language learner.
- Economically Disadvantaged: SES = Identified as economically disadvantaged, based on eligibility for K-12 federal assistance programs.
- Disability Status: DS = Identified as having a disability. Type of disability unspecified.

2.2 Scoring Model

One of the problems in the application of conventional pretrained LLMs, such as BERT (Devlin et al., 2019), is that the transformer architecture imposes a fixed context length (Vaswani et al., 2017; Mayfield and Black, 2020). There is an extensive body of literature that has addressed this length limitation, e.g. Longformer (Beltagy et al., 2020), Transformer-XL (Dai et al., 2019), and XLNet (Yang et al., 2019). These innovations are particularly suited for AES systems, which require longer context lengths.

Among the longer-context models, XLNet performs particularly well on AES and argumentation annotation (Ormerod et al., 2023). The key feature of XLNet is its recurrent form of attention (Dai et al., 2019).

Automated scoring generally benefits from using a regression head (with MSE loss) as opposed to a classification head (with cross-entropy loss) since regression parsimoniously retains the ordinal nature of score points (Ormerod et al., 2021).

We used the Adam optimizer with a weight decay mechanism (Loshchilov and Hutter, 2019). The learning rate was set to 5×10^{-6} with a linear learning rate scheduler, in batches of 8. Models were trained over 20 epochs, with early stopping determined by best performance on the dev set. To prevent out of memory errors, max token length was set to 2,048.

2.3 Performance Metrics

We assess the system's performance using the three standard metrics proposed by Williamson et al. (2012) for the evaluation of automated scoring systems. These include Cohen's quadratic weighted kappa (QWK, Cohen, 1960), standardized mean difference (SMD), and exact agreement.

These agreement statistics quantify the proximity of automated scores to human-assigned scores. Most operational standards consider model performance relative to human-human levels of agreement; however, only final score was included in the corpus. Nevertheless, Crossley et al. (2022) report that all essays were scored independently by two human raters and, across all PERSUADE items, the QWK was .745. Item-specific QWKs were not reported. SMD was also not reported. In the absence of double-scored data, a QWK of at least 0.7 and an SMD of at most 0.15 are commonly-accepted guidelines for adequate performance.

2.4 Analytic Approach Toward Bias

There are marginal (i.e. first-order) differences in score point distributions and in expected scores between demographic groups (Appendix A). For instance, female students generally score higher than male students on persuasive writing. It is pos-

[1]The PERSUADE corpus is publicly available online at https://github.com/scrosseye/persuade_corpus_2.0.

Prompt	Prompt Name	Grade	N_{Train}	N_{Dev}	N_{Test}
1	Phones and driving	N/A	558	140	464
2	Exploring Venus	10	740	185	923
3	Community service	8	608	153	773
4	Seeking multiple opinions	8	1232	309	7
5	Facial action coding system	10	880	221	1062
6	Distance learning	9-12	1192	299	656
7	Summer projects	9-12	696	175	872
8	Cell phones at school	8	663	166	824
9	Car-free cities	10	784	197	973
10	Grades for extracurricular activities	8	648	163	808
11	The face on Mars	8	654	164	764
12	Does the electoral college work?	9	1448	362	228
13	Driverless cars	10	1098	275	496
14	Mandatory extracurricular activities	8	668	167	824
15	"A Cowboy Who Rode the Waves"	6	546	137	682
Overall		**6-12**	**12422**	**3106**	**10356**

Table 1: A summary of how the data was split for training purposes.

sible that these group differences reflect biases in human-assigned scores; however, it is also possible that these group differences reflect legitimate differences in writing proficiency. Without additional information (e.g. a set of "unbiased" items, as would be used in an analysis of differential item functioning), the source of these differences cannot be determined.

The ambiguity of interpreting group differences extends to interpreting differences between automated and human-assigned scores. In *absolute* terms, for instance, differences could indicate that LLMs are introducing biases or, on the contrary, eliminating biases. As such, we limit ourselves to making claims in *relative* terms, i.e., do LLMs introduce biases relative to human scores?

2.5 Matching

On average, some groups scored higher or lower than others (e.g. female students scored higher than males, on average). To adjust for these marginal differences, we compared male and female students who received 1s to each other, male and female students who received 2s, etc., which is known as *exact matching* (Ho et al., 2011). Exact matching is ideal in this research context given that our sample is large, leaving very few students unmatched, even within specific prompts. As opposed to literally matching one student with another, we employ exact matching to produce a set of sample weights which, when taken as a whole, eliminate marginal group differences. These sample weights are used in subsequent analyses.

2.6 Group Difference Estimation

To compute human-XLNet scoring differences (i.e., relative bias), we estimated pairwise group differences. Regression estimates were produced using cluster-robust standard errors (Bell and McCaffrey, 2002; Pustejovsky and Tipton, 2018), as implemented by Blair et al. (2024) in R 4.3.1 (R Core Team, 2023). We used exact matching weights, described above, in these analyses.

2.7 Controlling False Discovery Rate

To avoid making spurious claims that are a product of random chance, we controlled the false discovery rate using the Benjamini-Hochberg (B-H) technique (Benjamini and Hochberg, 1995). We use the term *statistically significant* when an estimated p-value is below the B-H adjusted p-value. In practical terms, B-H adjusted p-values place an upper bound of .025 on "the probability of being erroneously confident about the direction of the population comparison" (Williams et al., 1999, p. 49).

2.8 Predicting Demographic Affiliation

We predict demographic affiliation using two complementary methods. The first, more conventional

method, is to train separate XLNet models to classify students' demographic affiliation based on their text responses. For example, we trained one model to predict gender, another model to predict race / ethnicity, etc.

The second method of predicting demographic affiliation was to use the hidden state from the scoring model for predictions. That is, for each demographic characteristic, linear models were trained using the hidden state as features.[2] More technically, the XLNet model used for scoring, $\mathcal{M}$, is a function of the input text, x, and can be broken into five distinct components:

$$\mathcal{M}(x) = \underbrace{(\sigma \circ \mathcal{L})}_{\text{Classifier}} \circ \underbrace{(\mathcal{S} \circ \mathcal{T} \circ \mathcal{E})}_{\substack{\text{feature} \\ \text{model}}}(x) \qquad (1)$$

where $\mathcal{E}$ is the embedding, $\mathcal{T}$ is the function for the layers of (recurrent) transformers, $\mathcal{S}$ is a summary layer that extracts the information for classification, $\mathcal{L}$ is a linear layer, and σ is the activation function. Conceptually, these five components can be grouped into a *feature model* and a *classifier*. The feature model maps text to a vector space of features that are subsequently used by the linear classifier to determine the score.

In predicting demographic characteristics, we used the following model:

$$\tilde{\mathcal{M}}(x) = (\sigma \circ \tilde{\mathcal{L}}) \circ (\mathcal{S} \circ \mathcal{T} \circ \mathcal{E})(x) \qquad (2)$$

Here, the feature model is frozen and $\tilde{\mathcal{L}}$ is optimized to predict demographic affiliation. If $\tilde{\mathcal{M}}$ can accurately distinguish demographic affiliation, using the language of Ormerod (2022), we say that the feature is *implicit* in the model. For example, in the ASAP dataset (Shermis, 2014), Ormerod (2022) demonstrated that essay length was an implicit feature of the model because it was a linear combination of the scoring features.

3 Results

We organize our findings around three foci. First, we evaluate the performance of XLNet to ensure it meets operational standards. Second, we assess the fairness of XLNet's automated scores by determining if there are any discrepancies, based on students' demographic affiliations, as compared to human-assigned scores. Finally, we determine the extent to which the scoring model has demographic features embedded within it.

3.1 Model Performance

We determined model performance on a prompt-by-prompt basis, as well as aggregated over all prompts. Table 2 summarizes the performance of the model in terms of three common agreement statistics: quadratic weighted kappa (QWK), standardized mean difference (SMD), and accuracy (all of which are described in greater detail in section 2.3).

Prompt	QWK	SMD	Acc	N
1	0.781	-0.066	0.683	464
2	0.856	0.003	0.677	923
3	0.800	-0.109	0.693	773
4	0.674	-0.312	0.429	7
5	0.865	-0.116	0.696	1062
6	0.875	0.042	0.697	656
7	0.813	-0.051	0.634	872
8	0.800	-0.021	0.717	824
9	0.796	-0.087	0.616	973
10	0.779	-0.025	0.699	808
11	0.818	0.063	0.658	764
12	0.863	-0.011	0.649	228
13	0.774	0.215	0.621	496
14	0.815	0.163	0.659	824
15	0.755	-0.040	0.691	682
Overall	0.864	-0.010	0.672	10356

Table 2: The performance of the model trained to the holistic scores in terms of the agreement with the human assigned scores.

Based on commonly-accepted operational standards, three items are in violation of these standards. More specifically, Prompts 4, 13, and 14 have high SMDs. Results for one of these items (Prompt 4), however, is unreliable due to the small test set sample size. Overall, however, XLNet performs well; indeed, in terms of overall QWK, XLNet exceeds human-human reliability.

3.2 Automated Scoring Biases

To measure automated scoring biases, we estimated pairwise differences between reference and focal groups. Table 3 displays the results of our automated scoring bias analysis, with standard errors in

[2]To clarify, XLNet (with a regression head) was first fine-tuned to predict score; after fine-tuning, we replaced the regression head with a classification head, froze all other layers, and fine-tuned again (using the same hyperparameters) to predict demographic characteristics.

Prompt	F-M	B-W	L-W	A-W	SES	ELL	DS
1	0.07 (0.06)	0.03 (0.04)	0.09 (0.05)	0.31 (0.22)			
2	0.10 (0.05)	0.00 (0.04)	-0.01 (0.05)	-0.06 (0.07)	-0.04 (0.05)	-0.10 (0.07)	-0.19 (0.02)
3	0.09 (0.04)	-0.15 (0.09)	-0.17 (0.03)	0.19 (0.09)	-0.11 (0.04)	-0.15 (0.07)	-0.35 (0.05)
5	0.07 (0.03)	-0.12 (0.02)	-0.12 (0.04)	0.01 (0.09)	-0.06 (0.01)	-0.09 (0.04)	-0.08 (0.03)
6	0.05 (0.02)	0.03 (0.10)	-0.08 (0.07)	0.07 (0.13)	-0.15 (0.08)	-0.28 (0.10)	-0.07 (0.08)
7	0.04 (0.04)	-0.29 (0.06)	-0.12 (0.04)	0.10 (0.04)	-0.06 (0.02)	-0.13 (0.09)	-0.12 (0.06)
8	0.09 (0.04)	-0.19 (0.07)	-0.11 (0.02)	0.14 (0.16)	-0.12 (0.03)	-0.19 (0.03)	-0.18 (0.10)
9	0.12 (0.02)	-0.13 (0.05)	-0.06 (0.05)	0.16 (0.17)		-0.16 (0.08)	
10	0.13 (0.03)	-0.20 (0.09)	-0.14 (0.06)	-0.04 (0.06)	-0.19 (0.05)	-0.26 (0.11)	0.03 (0.11)
11	0.09 (0.06)	-0.08 (0.04)	-0.02 (0.01)	-0.06 (0.05)	-0.12 (0.07)	-0.33 (0.09)	-0.11 (0.06)
12	0.07 (0.08)						
13	0.14 (0.04)	-0.08 (0.03)	-0.02 (0.04)	0.05 (0.17)	-0.27 (0.05)	-0.37 (0.33)	0.04 (0.19)
14	0.12 (0.06)	-0.12 (0.06)	-0.09 (0.02)	0.04 (0.01)	-0.17 (0.03)	-0.15 (0.05)	-0.09 (0.05)
15	0.04 (0.03)	-0.08 (0.07)	0.00 (0.08)	0.11 (0.10)	-0.13 (0.07)	-0.31 (0.09)	0.08 (0.14)
Overall	**0.06 (0.01)**	**-0.07 (0.01)**	-0.06 (0.02)	0.07 (0.02)	-0.10 (0.02)	-0.10 (0.04)	-0.07 (0.02)

Table 3: Biases in XLNet scores, relative to human-assigned scores. Pairwise group differences are presented as z-scores. Bold font indicates statistically significant differences.

parentheses. Score differences were normalized so that units are in standard deviations (i.e. they may be interpreted as z scores). More specifically, a difference of 0 indicates that there was no difference between focal and reference groups; a negative difference indicates that the focal group received a lower score, on average, compared to the reference group; and a positive difference indicates that the focal group received a higher score. Differences that were statistically significant are presented in bold.

Group differences varied across prompts, but trends were generally consistent. We found no statistically significant group differences within specifics prompts.

Overall, however, we found that XLNet gave higher scores to female students compared to male students ($z = 0.06$, $SE = 0.01$, $p = .0012$), and lower scores to Black students compared to White students ($z = -0.07$, $SE = 0.01$, $p = .0023$). These differences are consistent with marginal differences observed between these groups, based on human-rater scores (Table 5). That is, XLNet magnified marginal between-group differences; the effect size, however, was small. Students with low SES status and English Language Learner status also scored lower than their respective reference groups; these differences, however, were not statistically significant.

3.3 Model-Embedded Demographics

To determine if demographic information was embedded within the scoring model, we predicted demographic affiliation from the hidden state of the model. The right side of Table 4 ("Score Features") presents the results of these analyses, with QWK (or κ) as the effect size.

According to McHugh (2012), a κ value within the range of $0 \leq \kappa \leq 0.2$ is considered to have "no agreement," $0.2 < \kappa \leq 0.4$ is considered "minimal," $0.4 \leq \kappa \leq 0.6$ is "moderate," $0.6 < \kappa \leq 0.8$ is "substantial," and anything above 0.8 is "almost perfect."

For nearly all prompts, effect sizes range from "no agreement" to "minimal agreement." The one exception is predicting ELL status in Prompt 6 ($\kappa = 0.75$), which is a substantial effect size. This suggests that XLNet was able to distinguish ELL status quite well based on students' essay responses for this prompt.

In interpreting these results, it is important to bear in mind that we have not controlled for marginal differences in students' scores or factors associated with students' scores. Some of these additional factors are listed in Appendix A. For example, length is associated with students' scores and it is well-documented that female students tend to write more than males. When essay length is used to predict gender, the strength of the relationship is $\kappa = 0.058$. Note that this effect size is only slightly better than randomly guessing the gender of the student. Using the average word count, word-length, number of sentences, and Flesch–Kincaid as features to determine gender, we obtained a κ statistic of 0.106, and $\kappa < 0.06$ for all races / ethnicities, disability status, and ELL status.

We not only predicted demographic affiliation from the scoring model, but also trained separate XLNet models to predict demographic affiliation

directly from students' essays. The left side of Table 4 ("Text") presents these results. These results serve as a useful comparison, since they serve as an upper-bound of how well XLNet can predict student groups based on essay responses. κ values seem particularly high for SES and ELL.

4 Discussion

4.1 Conclusions

This study makes an important contribution to the growing body of research on bias in AES systems based on LLMs. Although XLNet generally demonstrated strong performance on key metrics compared to human raters, it also magnified marginal differences between groups, relative to human-assigned scores. In particular, relative to human-assigned scores, XLNet was found to be more generous to female students compared to male students and White students compared to Black students. Additionally, we found evidence that these group differences were embedded in the hidden layer of the model.

Although effect sizes of biases were small, in large-scale assessments even small differences can affect many students. Furthermore, in high stake settings (e.g. high-school exit exams), such differences can result in failure to meet graduation requirements. XLNet magnified marginal differences, a finding consistent with other research (Kwako et al., 2023); this indicates that marginalized populations may be particularly at risk of unfair scoring.

Overall, this study demonstrates the importance and feasibility of comprehensive bias evaluations when deploying AI scoring in high-stakes educational settings. Responsible use of automated systems requires evidence that they do not create or worsen inequities for marginalized student populations. With careful design and monitoring, LLMs should help make writing assessment more consistent, reliable, and constructive for all students.

4.2 Limitations

As stated above (Section 2.4), our claims are limited to evaluating biases relative to human scores. Yet human scores themselves are often biased (e.g. Zechner, 2019). Thus, it is possible that XLNet is more fair than human raters, in spite of it magnifying marginal group differences relative to human raters. Differential item functioning (DIF, Angoff, 1993) accounts for these potential biases by relying on an "unbiased" set of anchor items. The PERSUADE corpus does not include such data, however, and there is no public dataset currently available that would permit DIF analyses.

Results showed that demographic affiliations were embedded in the hidden layer of the XLNet scoring model. Yet, without further investigation, we are unable to determine if this information is used (e.g. as an implicit feature) in generating students' essay scores.

Lastly, we recognize that this study was limited to analyzing biases within a single LLM model and dataset. Further research could evaluate other state-of-the-art models and diverse essay sets to determine the extent to which findings generalize.

4.3 Further Research

The limitations of this study, noted above, reveal several promising paths forward. There is room, for instance, to explore additional LLM models (beyond XLNet) and additional datasets. It would also be valuable to investigate sources of group differences (e.g. language differences between groups), and to determine if these group differences are construct relevant or not. Construct (ir)relevance is important to consider, as it affects which debiasing strategies would be viable (Kwako, 2023).

Along the lines of debiasing, it would be helpful to explore bias mitigation techniques at both the training and scoring stages. For example, if demographic affiliation is an implicit feature (i.e. $\mathcal{L}(x) = \alpha x + \beta$, and $\tilde{\mathcal{L}}(x) = \tilde{\alpha} x + \tilde{\beta}$), then we could potentially use orthogonal projection to optimize α on the vector-subspace orthogonal to $\tilde{\alpha}$. This might mitigate the effect of any features the model is using to distinguish demographic information. This may, however, come at some cost to model performance.

References

William H Angoff. 1993. Perspectives on differential item functioning methodology. In *Differential item functioning*, pages 3–23. Routledge.

Robert M Bell and Daniel F McCaffrey. 2002. Bias reduction in standard errors for linear regression with multi-stage samples. *Survey Methodology*, 28(2):169–182.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The Long-Document Transformer. ArXiv:2004.05150 [cs].

#	Text							Score Features							
	G	W	L	B	SES	ELL	DS	G	W	L	B	A	SES	ELL	DS
1	0.16	0.25	0.01	0.21				0.07	0.08	0.05	0.14	-0.01			
2	0.22	0.36	0.25	0.07	0.31	0.16	0.13	0.19	0.18	0.09	0.09	0.16	0.17	0.33	0.28
3	0.23	0.26	0.24	0.11	0.30	-0.00	0.00	0.24	0.29	0.16	0.04	0.07	0.18	0.09	0.20
4	-0.08	0.05	0.00	0.22	0.28	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
5	0.26	0.35	0.24	0.10	0.33	0.42	0.05	0.26	0.26	0.18	0.21	0.14	0.28	0.28	0.23
6	0.17	0.43	0.45	0.10	0.46	0.77	0.02	0.24	0.34	0.38	0.14	0.06	0.28	0.75	0.15
7	0.27	0.18	0.10	0.16	0.17	0.29	0.00	0.22	0.14	0.08	0.18	0.14	0.19	0.21	0.07
8	0.23	0.27	0.24	0.09	0.34	0.30	0.00	0.25	0.25	0.16	0.13	0.05	0.21	0.13	0.13
9	0.26	0.29	0.18	0.12		0.06		0.27	0.24	0.10	0.11	0.07		0.23	
10	0.26	0.25	0.20	0.12	0.32	0.21	0.00	0.19	0.28	0.11	0.13	0.06	0.20	0.15	0.05
11	0.28	0.20	0.06	0.15	0.23	0.00	0.00	0.21	0.14	0.00	0.21	0.00	0.18	0.00	0.00
12	0.32	0.17	0.16	0.00				0.21	0.15	0.08	0.00	-0.05			
13	0.31	0.12	0.08	0.12	0.04	0.00	0.00	0.20	0.05	0.07	0.14	0.13	0.07	0.00	0.00
14	0.28	0.18	0.21	0.08	0.37	0.35	0.01	0.28	0.12	0.08	0.08	0.14	0.26	0.28	0.10
15	0.25	0.21	0.11	0.16	0.19	0.00	0.00	0.06	0.11	0.06	0.09	0.04	0.09	0.14	0.04

Table 4: The ability of our models to determine demographic affiliation measured by Cohen's kappa statistic. The columns under Text present κ values for language models trained on the text, while the columns under Score Features are linear models whose features coincide with those used to determine score.

Yoav Benjamini and Yosef Hochberg. 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.2517-6161.1995.tb02031.x.

Rishabh Bhardwaj, Navonil Majumder, and Soujanya Poria. 2021. Investigating Gender Bias in BERT. *Cognitive Computation*, 13(4):1008–1018.

Graeme Blair, Jasper Cooper, Alexander Coppock, Macartan Humphreys, and Luke Sonnet. 2024. *estimatr: Fast Estimators for Design-Based Inference*. R package version 1.0.2.

Jacob Cohen. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37–46. Publisher: SAGE Publications Inc.

Scott A. Crossley, Perpetual Baffour, Yu Tian, Aigner Picou, Meg Benner, and Ulrich Boser. 2022. The persuasive essays for rating, selecting, and understanding argumentative and discourse elements (PERSUADE) corpus 1.0. *Assessing Writing*, 54:100667.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc V. Le, and Ruslan Salakhutdinov. 2019. Transformer-XL: Attentive Language Models Beyond a Fixed-Length Context. ArXiv:1901.02860 [cs, stat].

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Technical Report arXiv:1810.04805, arXiv. ArXiv:1810.04805 [cs] type: article.

Daniel E. Ho, Kosuke Imai, Gary King, and Elizabeth A. Stuart. 2011. MatchIt: Nonparametric preprocessing for parametric causal inference. *Journal of Statistical Software*, 42(8):1–28.

Alexander Kwako, Yixin Wan, Jieyu Zhao, Mark Hansen, Kai-Wei Chang, and Li Cai. 2023. Does bert exacerbate gender or l1 biases in automated english speaking assessment? In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 668–681.

Alexander James Kwako. 2023. *Mitigating Gender and L1 Biases in Automated English Speaking Assessment*. Ph.D. thesis, University of California, Los Angeles.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. ArXiv:1711.05101 [cs, math].

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.

Mary L. McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia Medica*, 22(3):276–282.

Christopher Ormerod, Amy Burkhardt, Mackenzie Young, and Sue Lottridge. 2023. Argumentation Element Annotation Modeling using XLNet. ArXiv:2311.06239 [cs].

Christopher Ormerod, Susan Lottridge, Amy E. Harris, Milan Patel, Paul van Wamelen, Balaji Kodeswaran, Sharon Woolf, and Mackenzie Young. 2022. Automated Short Answer Scoring Using an Ensemble of Neural Networks and Latent Semantic Analysis Classifiers. *International Journal of Artificial Intelligence in Education*.

Christopher M. Ormerod, Akanksha Malhotra, and Amir Jafari. 2021. Automated essay scoring using efficient transformer-based language models. Number: arXiv:2102.13136 arXiv:2102.13136 [cs].

Christopher Michael Ormerod. 2022. Mapping Between Hidden States and Features to Validate Automated Essay Scoring Using DeBERTa Models. *Psychological Test and Assessment Modeling*, 64(4):495–526.

James E. Pustejovsky and Elizabeth Tipton. 2018. Small-Sample Methods for Cluster-Robust Variance Estimation and Hypothesis Testing in Fixed Effects Models. *Journal of Business & Economic Statistics*, 36(4):672–683. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/07350015.2016.1247004.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Pedro Uria Rodriguez, Amir Jafari, and Christopher M. Ormerod. 2019. Language models and Automated Essay Scoring. Number: arXiv:1909.09482 arXiv:1909.09482 [cs, stat].

Mark D. Shermis. 2014. State-of-the-art automated essay scoring: Competition, results, and future directions from a United States demonstration. *Assessing Writing*, 20:53–76.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Valerie SL Williams, Lyle V Jones, and John W Tukey. 1999. Controlling error in multiple comparisons, with examples from state-to-state differences in educational achievement. *Journal of educational and behavioral statistics*, 24(1):42–69.

David M. Williamson, Xiaoming Xi, and F. Jay Breyer. 2012. A Framework for Evaluation and Use of Automated Scoring. *Educational Measurement: Issues and Practice*, 31(1):2–13. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1745-3992.2011.00223.x.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. XLNet: Generalized Autoregressive Pretraining for Language Understanding. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Klaus Zechner. 2019. Summary and outlook on automated speech scoring. In *Automated speaking assessment*, pages 192–204. Routledge.

A Differences across student groups

This appendix reports descriptive statistics of essays written by students, disaggregated by demographic affiliations. In addition to known discrepancies between the lengths of essays between certain groups (notably male and female students), we present the average word length, number of sentences, and the Flesch-Kincaid grade, which is a common readability measure defined by

$$G = \alpha\left(\frac{\text{total words}}{\text{total sentences}}\right) + \beta\left(\frac{\text{total syllables}}{\text{total words}} + \gamma\right) + \gamma \quad (3)$$

where $\alpha = 0.39$, $\beta = 11.8$, and $\gamma = 15.59$. These statistical differences in essay texts, by demographic affiliations, are presented in Table 5.

Category	Subgroup	Rep.	Averages Score	Words	Word Len.	Sent.	F.K.
Gender	Male	49.5%	3.20	404	4.40	19.3	9.32
	Female	50.5%	3.43	432	4.45	21.9	8.70
Race/ Ethnicity	White	44.5%	3.42	427	4.41	21.4	8.60
	Hispanic/Latino	25.2 %	3.08	398	4.40	19.0	9.50
	Black/African American	19.1%	3.12	393	4.43	19.3	9.26
	Asian/Pacific Islander	6.7%	3.37	504	4.59	25.1	9.22
	Two or More	3.9%	3.45	429	4.46	21.1	8.87
	Native American	0.5%	3.02	369	4.35	19.3	8.31
ELL	Identified	8.6 %	2.69	374	4.42	16.5	10.7
	Not Identified	86.4%	3.35	421	4.42	20.9	8.87
Economic Disadvantage	Identified	37.1 %	2.98	367	4.36	18.0	9.19
	Not Identified	42.8%	3.65	446	4.44	22.0	8.9
Disability Status	Identified	10.3%	2.72	360	4.36	17.0	9.6
	Not Identified	69.8%	3.33	416	4.41	20.6	8.95

Table 5: Some key statistical differences between the nature of the scores and essays, disaggregated by demographic affiliation.