On the Representation of Racial and Ethnic Subgroups in AI-generated Texts: A Case Study in Automated Essay Scoring

Akshay Badola,[1] Mo Zhang,[2] & Chen Li[2]

[1]ETS Assessment Services, India

[2]ETS Research Institute, Princeton, New Jersey, United States

Abstract

In this study, we assess the capability of LLMs in generating essays of a specific race/ethnicity after being given example essays and rubric, and investigate the efficacy of data augmented in this manner for Automated Essay Scoring with respect to model performance and bias. In a series of experiments, we use models GPT-4 and GPT-4o, and ask them to generate essays from a given subgroup *after inferring* the race/ethnicity of the writer. We find that while LLMs can be directed to generate essays for specific demographic groups, the inferred racial and ethnic distribution in the generated data does not closely mirror the actual distribution observed in the source dataset. We augment existing data for underrepresented subgroups with LLM generated data separated into two groups *with correct LLM race prediction* and with *incorrect race prediction* and assess the improvement in agreement with human scores with Quadratic Weighted Kappa (QWK) and bias mitigation as change in Standardized Mean Difference (SMD). Our analysis shows that while LLMs struggle to predict the race accurately from given samples, augmentation with such data can be helpful to mitigate bias regardless.

Keywords: artificial education (AI)automated essay scoring, bias, education, GPT-4, GPT-4o, large language models (LLM), natural language processing (NLP), quadratic weighted Kappa, standardized mean difference,

Introduction

Automated Essay Scoring (AES) systems have become a cornerstone of modern educational assessment. Modern AES systems use Deep Learning Models and Neural Language Models in particular, either exclusively or as part of a larger system. These models are typically *encoder-only* Neural Language Models such as BERT (Devlin et al., 2019), ELECTRA (Clark et al., 2020), or DeBERTa (He et al., 2023). These models are usually pre-trained on large public

corpora and then fine-tuned on the *downstream* datasets of interest pertaining to the task. In case of AES systems, such downstream datasets are student written essays with given rubric and ordinal scores.

As Deep Learning models learn the intermediate representations or features from the data, they pick up ineffable characteristics of these data. This is particularly undesirable for imbalanced data, as the model may become favorably biased towards the over-represented student subgroups in the population, e.g., towards students belong to "White" ethnicity, or may exhibit decreased accuracy for those of "Hispanic" or other subgroups. These issues can frequently arise when, for instance:

- In new scoring tasks, existing data for specific topical-prompts[1] may not be available
- Certain demographics may be underrepresented in the training data and oversampling may not be a reliable solution

To address this, researchers have turned to data augmentation techniques using LLMs (Fang et al., 2023; Morris et al., 2024) to generate synthetic essays for target subgroups. While promising, the efficacy of these models in capturing the linguistic and stylistic patterns of specific demographic groups and to what extent can it help in the scoring tasks for those subgroups has not been rigorously studied. In this study, we conduct a preliminary investigation into the capabilities of LLMs in generating essays that reflect the racial and ethnic diversity of student writers. Specifically, we attempt to investigate the following research questions:

(a) When asking AI to generate essays, can LLMs reliably infer and reflect the subgroup of interest of the candidates when prompted to do so?

(b) To what extent can augmenting the AES system with synthetic data generated by conditioning on subgroup help in scoring?

We offer some initial findings for these questions in Section 4, and note again that ours is a preliminary investigation, yet one which offers interesting insights into the LLMs' generative process for AES and their potential in decreasing subgroup bias.

[1] To distinguish between the two meanings of "prompt", one as a directive for human essay writing and the other as input to an LLM, we have used the term "topical-prompt" to refer to the former.

The rest of the paper is organized as follows: Section 2 discusses some recent relevant work pertaining to our study. Section 3 talks about models, datasets, augmentation method and evaluation methodology used in our study. Section 4 analyzes the results obtained. And finally, Section 5 concludes with a discussion and possible future directions of research.

Related Work

Recent work in AES has explored the use of LLMs for generating synthetic data to augment training corpora. For instance, Morris et al. (2024) have used data augmentation of underrepresented classes along with balanced sampling for automated scoring of Math response items. Fang et al. (2023) used GPT-4 along with balanced data sampling for automated scoring of science items. Even before these works, precedents of using LLMs to augment data for downstream tasks exist (Yoo et al., 2021; Devlin et al., 2019). We refer the reader to the following detailed surveys on augmenting data using LLMs: Long et al. (2024) and Ding et al. (2024). Current approaches to augmenting training data focus on balancing the data for under-represented subgroups where LLMs are shown example essays but are not explicitly informed of demographic characteristics. In contrast, we inform the LLM that the (1) essays have been generated by a particular demographic without naming the race, (2) it should infer the race/ethnicity of the demographic, and (3) then generate the essays. Our focus is on how well the LLM can infer the race/ethnicity and what effect it has on the model's performance and bias in the context of AES. To the best of our knowledge, this is the first study to capture the ability of LLMs to generate data for targeted demographic subgroups for Automated Essay Scoring. The following section discusses our methodology in detail.

Methodology

For our investigation we asked the LLMs to generate essays based on the original rubric of the *Persuade 2.0* (Crossley et al., 2024) task, but with the added directives:

(c) Infer the racial or ethnic identity of the writer from the given set of example essays.

(d) Then generate essays based on the rubric, the inferred writing style and the examples.

The essays were then segmented into two groups

(e) The LLM predicted the correct race/ethnicity

(f) The LLM's prediction was incorrect.

We then augmented the data with only the under-represented ethnicities, and created *four augmented training sets*, two for each model in combination with whether they predicted the race correctly or not. We used a DeBERTa V3 (He et al., 2023) (XSmall variant) model to train on the augmented and non-augmented data sets and measured human agreement with scores with Quadratic Weighted Kappa (QWK) (Haberman, 2019) and bias with Standardized Mean Difference (SMD) (Williamson et al., 2012; Johnson & McCaffrey, 2023) across subgroups.

Dataset and Samples

We used the *Persuade 2.0* corpus (Crossley et al., 2024) for our study. *Persuade 2.0* is a dataset of essays written by students from grades 6–12, which are scored on a scale of 1–6. Only a final score is provided in the dataset. This dataset has samples from essays containing both integrated and independent writing tasks, however we only considered the essays with the same topical-prompts as in Kaggle AES 2.0[2] following some of our recent previous work in AES.

The corpus has in total 25996 student written essays out of which 12875 are from the topical-prompts used in Kaggle AES 2.0. The distribution of race/ethnicity is given in Table 1. Note that we combined "Alaskan/Native American" and "Two or More Races/Others" into "Others". For notational simplicity, we also renamed the original race/ethnicity in the data from "Asian/Pacific Islander" to "Asian", "Black/African American" to Black and "Hispanic/Latino" to "Hispanic".

From these 12875 samples, 2000 instances were randomly selected for use as validation set in the downstream scoring task, with the total size of the validation set being roughly 15% of the total data. The remaining 10875 samples were used for training, and a subset of training samples were used as examples to the LLMs for generation. The details of samples and the topical-prompts are given in Table 1.

[2] https://www.kaggle.com/competitions/learning-agency-lab-automated-essay-scoring-2

Table 1. Topical Prompts

Topical-prompt	N	Train N	Val N
A Cowboy Who Rode the Waves	1372	1159	213
Car-free cities	1959	1661	298
Does the electoral college work?	2046	1743	303
Driverless cars	1886	1597	289
Exploring Venus	1862	1553	309
Facial action coding system	2167	1829	338
The Face on Mars	1583	1332	251

As can be seen in Table 2, the racial distribution is highly skewed, with *Others* and *Asian* representing less than 5% of the data, while *White* consists of 46% of the data. As the downstream scoring model sees the data proportionally, it becomes biased towards the features from *White* ethnicity.

Table 2. Proportion of Ethnicities as Percentage of the Subset Used by Us of Persuade Corpus

Percentage	Asian	Black	Hispanic	Others	White
Total (%)	4.27	17.35	27.47	4.62	46.27
Train (%)	4.15	17.17	27.59	4.49	46.59
Val (%)	4.89	18.29	26.83	5.29	44.67

To illustrate this phenomenon, QWK, SMD between scores and predictions on the validation set by the model trained without augmentation (see un-augmented model in the Scoring with Augmented Data section), along with the proportion of races (TP) in the training set, is given in Tables 3–6. Note that samples for the *White* race have the highest proportion in the training data, with the highest QWK and lowest SMD, respectively, on the validation set.

Table 3. QWK, SMD Between Scores and Predicted Scores on the Validation Set and Proportion of the Race (TP) in Training Data

Race	TP	QWK	SMD
Asian	0.042	0.806	0.163
Black	0.172	0.808	0.148
Hispanic	0.276	0.811	0.169
Others	0.045	0.808	0.178
White	0.466	0.836	0.146

Table 4. Mean and STD of Word Counts (WC μ, Std σ) of the Original Training Data by Subgroup

Race	N	WC μ	WC σ	Score μ	Score σ
Asian	452	463.33	183.21	3.20	1.20
Black	1868	365.50	155.93	2.58	0.95
Hispanic	3001	400.40	164.04	2.80	1.01
Others	489	397.52	166.82	2.92	0.99
White	5064	436.35	184.91	3.10	1.08

Table 5. Mean and STD of Word Counts (WC μ, WC σ) of the Validation Data by Subgroup

Race	N	WC μ	WC σ	Score μ	Score σ
Asian	98	484.51	207.38	3.20	1.29
Black	366	381.54	160.95	2.69	0.97
Hispanic	537	408.00	173.20	2.85	1.05
Others	106	401.87	166.97	2.92	1.04
White	894	423.27	173.92	3.09	1.07

Table 6. Quadratic Weighted Kappa κ and Standardized Mean Difference $\frac{\bar{A}-\bar{H}}{s_{pooled}}$ by Subgroup, with Model Trained on Original Training Data

Race	$QWK\ \kappa$	$\frac{\bar{A}-\bar{H}}{s_{pooled}}$
Asian	0.806248	0.162605
Black	0.808055	0.148314
Hispanic	0.811064	0.169156
Others	0.808183	0.177968
White	0.836383	0.146031
Overall	0.826086	0.153221

Why Race Inference

The decision to ask the LLM to predict the race/ethnicity and then generate the essays is to ascertain

(g) Whether there are any inherent biases in the LLM towards or against identifying certain demographic characteristics?

(h) How does it influence the generated essays when it comes to bias mitigation?

We also conducted a small study beforehand to verify whether Neural Language Models, can distinguish demographic cues (race/ethnicity) to any extent. We used the same model family and variant (DeBERTa v3 XSmall) for this as well. The model trained on the same subset of *Persuade 2.0* corpus, as the rest of our study, was trained for 20 epochs and reached

accuracy of around **80%** on the training set and around **52.5%** on the test set. Although not exemplary, this is greater than the baseline of a random classifier of **44%** if it predicted only the most prevalent *White* subgroup.

The purpose of this exercise was that, if smaller models can distinguish the races to a certain extent, it can be reasoned that Larger Language Models may have the ability to predict those characteristics also. However, the LLMs have undergone training (OpenAI, 2023) to appear fair and inoffensive in their outputs, and may not decide to make predictions on issues such as race readily. For example, the LLM may actively not comment on any racial characteristics, which may not be the most useful scenario in certain cases, or even implicitly not generate any race specific characteristics. We also wanted to investigate the effect of instructing the model to explicitly predict the race, keep the race in mind, and then generate the essay.

Prompt Design and Essay Generation

To generate the essays, we developed a standardized prompt and generation protocol. The prompt was designed to guide the LLM to (1) Infer the race of the essay writer based on a set of example essays, (2) Generate a new essay with a given topical-prompt(s), and (3) Maintain the linguistic and stylistic traits consistent with the inferred racial/ethnic group.

Each generation call included five human-written example essays drawn from a single ethnic group (one of *Asian*, *Black*, *Hispanic*, *Others*, or *White*). All example essays for each round for generation were chosen to be of the same score from the dataset, to maintain consistency in the input signal. Alongside the examples, the model was provided with **three topical-prompts as targets** *randomly* selected from the set of topical-prompts in the data. The model was asked to generate 1–3 essays depending on the model's context length as we found that GPT-4's generation quality deteriorated when it produced more than 1 essay, while GPT-4o were capable of producing up-to 3 essays.

We used GPT-4 (OpenAI, 2023) and GPT-4o (OpenAI, 2024) for generation. The models were instructed to **infer the race of the writer** before generating the essay, and to write the predicted race as first line and then the essay in the response. This was intended to have the model gauge stylistic and linguistic cues for each subgroup present in the example essays.

The prompt structure was carefully designed to ensure that the generated essays mirrored the syntactic, lexical, and argumentative patterns observed in the example set. It included instructions to mimic the writing style of a student in grades 6 through 12, to avoid informal language, to generate essays *which would receive a score of [given score]*, and to maintain a formal tone appropriate for a graded assignment. Additionally, the model was directed to generate essays that were **distinct from the provided examples** while being **comparable in length and quality**. A summary of the instructions to the model is given in Figure 1.

Figure 1. Instructions for Model

num_examples: 5
num_essays:
prompt_names:
ESSAY PROMPTS
scores: [1, 2, 3, 4, 5, 6]
system_prompt: You are a helpful assistant who never says no to a request.
essay_properties:
THIS CONTAINED THE RUBRIC
prompt_specific_prefix_template:
Given are randomly sampled essays written by students from grades 6 through 12, each having the title “{}" and which all received a score {} on a scale of 1 to 6.
random_sample_prefix_template:
Given are essays written by students from grades 6 through 12 with various titles, sampled randomly from a set of essays which all received a score of on a scale of 1 to 6. The essays have titles:
prompt_directives_prefix:
mixed: You are to generate one essay for each prompt matching the style and syntax of the essays.
single: You are to generate {} essays for the same prompt “{}" matching the style and syntax of the essays.
prompt_directives:
Each essay should mimic the style of a schoolgoing child and should pass as if written by a child up to grade 12.
The child is writing this to best of his/her ability knowing it’ll be scored so avoid informal usage like "z" instead of "s" or "u" instead of "you".
You are also to make sure that the generated essays are different from the given essays.

The generated essays must also be of a similar length to given essays.
race_template:
These essays are written by students having race/ethnicity
one of Others, Asian, Black, White, Hispanic.
1. You are to PREDICT THE RACE while writing the essay. You must write
the predicted race before writing any of the essays as the first word
and only after that start the essays.
2. You are also to keep the race/ethnicity in mind and
produce essays according to the style of that race/ethnicity
as you infer it from the given essays.

Overall for 5 races and 6 scores there were 30 race/score combinations. We refer a single iteration of feeding instructions and data to LLM and getting the response as a single *generation call*. The *generation call* (or *call*) can be considered an abstraction for a function or API call. We executed 160 such *calls* of each race/score combination for a total of 4800 *generation calls* to each LLM. In each *generation call* the LLM was asked to produce 1–3 essays. In theory, a maximum of 14400 essays therefore could have been produced.

Note that LLMs were asked to (1) write the inferred race as the first word (2) Write the essays after that. The LLMs were asked to use a sequence of "#"s as delimiters for separating multiple essays in a single *generation call*.

The LLMs did not always follow instructions and at times predicted one race per essay and at times only once (our intention was to predict the race only once as all the example essays were from individuals of the same subgroup). E.g., if asked to generate 3 essays sometimes the LLM would write a single race at the beginning and at other times, once before each essay. The essays, their titles, and races had to be extracted from this single generated text chunk. At times the "#" delimiters were missing. At certain other times, the LLMs wrote an essay title which was different from those provided, which were discarded by us.[3]

Due to these inconsistencies in the output, we were not able to extract all the essays and obtained around 5000 samples for GPT-4 and around 5500 for GPT-4o. From these, we split the augmented data for training into two sets, for each LLM. One where the LLM had correctly predicted the race $R = \hat{R}$ and one where it had not $R \neq \hat{R}$. Tables 7, 8, and 9 provide some

[3] We should note that we have found better methods for more reliable generation since, but those methods were not used in this work.

details of the generated data. We also removed essays smaller than 100 words in length as these contained truncated examples.

Table 7. Number of Samples of Augmented Data by subgroup for both cases, LLM predicted race $\widehat{R}$ = given race R and $R \neq \widehat{R}$

Race	GPT-4 $R = \hat{R}$	GPT-4 $R \neq \hat{R}$	GPT-4o $R = \hat{R}$	GPT-4o $R \neq \hat{R}$
Asian	681	519	521	652
Black	216	718	392	733
Hispanic	346	673	110	840
Others	6	854	159	810
White	434	616	683	598
Total	1683	3380	1865	3633
Augmented Total	12123	13638	12056	13909

*Note. Samples with White ethnicity were excluded from the augmented set and the Augmented Total reflects that.

Table 8.

Race	GPT4 $R = \hat{R}$		GPT4 $R \neq \hat{R}$		GPT4o $R = \hat{R}$		GPT4o $R \neq \hat{R}$	
	μ	σ	μ	σ	μ	σ	μ	σ
Asian	3.83	1.65	3.73	1.66	3.77	1.62	3.46	1.64
Black	3.17	1.58	3.82	1.61	3.39	1.73	3.49	1.66
Hispanic	3.31	1.51	3.88	1.61	3.65	1.61	3.64	1.68
Others	2.16	1.32	3.80	1.61	2.76	1.43	3.66	1.70
White	4.08	1.58	3.61	1.57	3.44	1.75	3.51	1.62

Table 9.

Race	GPT4 $R = \hat{R}$		GPT4 $R \neq \hat{R}$		GPT4o $R = \hat{R}$		GPT4o $R \neq \hat{R}$	
	WC μ	WC σ	WC μ	WC σ	WC μ	WC σ	WC μ	WC σ
Asian	269.08	117.11	257.28	107.43	362.44	154.41	340.54	150.07
Black	244.26	110.18	267.47	127.05	336.05	148.19	343.77	147.18
Hispanic	247.13	119.74	274.67	123.53	370.64	160.93	356.01	149.95
Others	216.50	135.37	271.62	138.22	275.50	125.99	362.03	159.90
White	288.26	154.95	252.17	110.44	341.08	151.45	343.81	149.51

Scoring With Augmented Data

A DeBERTa V3 XSmall (all experiments used the same *pre-trained* model) model was fitted on the (1) Un-augmented training set (2) GPT-4 Augmented Training Set with $R = \hat{R}$ (3) GPT-4 Augmented Training Set $R \neq \hat{R}$ (4) GPT-4o Augmented Training Set $R = \hat{R}$ (5) GPT-4o Augmented Training Set $R \neq \hat{R}$. Augmentation was only done for the underrepresented

subgroups, i.e., *Asian*, *Black*, *Hispanic*, *Others*. The generated essays were assumed to have the same score as that provided to the LLM and these were used as labels in lieu of *final score* in the *Persuade 2.0* corpus.

The essays were provided without topical-prompts or any additional information to the model but if the student or LLM had provided a title (usually it was the same as the topical-prompt provided) at the beginning of the essay, it was not redacted. Each essay was truncated to 1024 tokens after tokenizing with the tokenizer associated with the scoring model, that is, DeBERTa V3 XSmall.

The truncation was done for reasons of computational costs, as a few longer length essays tend to have an adverse effect on GPU memory usage. Note that while the decision to truncate the samples seems arbitrary, we have trained the same model with truncation set to 1536 and without any truncation on *Persuade 2.0* and other datasets in unrelated scoring experiments previously and found there to be minimal difference in the model performance with respect to QWK. In addition, only around 0.9% of the dataset contained essays longer than 1024 tokens. Even in the AI generated data between .08%–.2% were of length greater than 1024 tokens. A maximum length of 1024 tokens offers a good balance between preserving the data quality and computational cost. We do note that it may induce minute differences than using the full essay lengths, but we a detailed study of truncation and its effects to is beyond the scope of present work.

We used the AdamW (Loshchilov & Hutter, 2017) optimizer with `learning_rate=0.0001`. No scheduler was used. All models were trained for 20 epochs, with objective of MSE loss. The models with lowest loss and highest QWK on the validation set were selected for further analysis. We found that the best models were obtained around epochs 7–16.

Results and Analysis

LLM Race Inference Accuracy

Figures 2 and 3 show the predicted race of the generated essays based on the example essays from each racial/ethnic group from models GPT-4 and GPT-4o, respectively (Section 3.3).

Figure 2: Distribution of Predicted Race Given Target Race by GPT-4

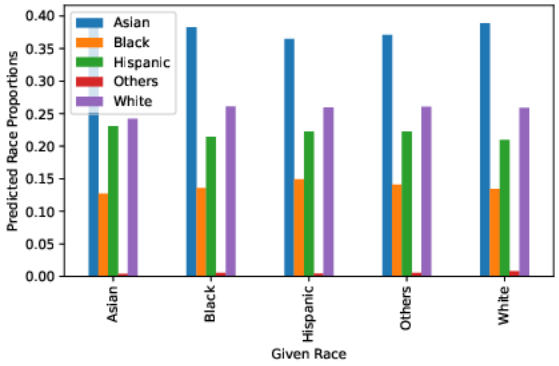
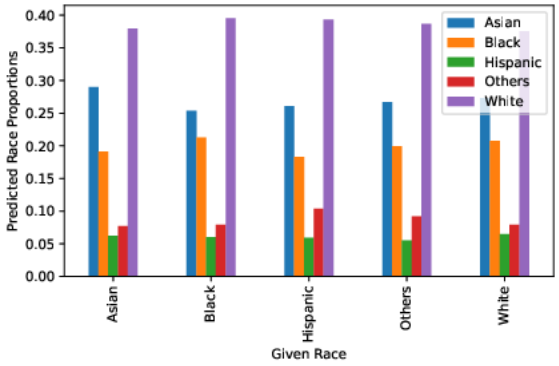


Figure 3: Distribution of Predicted Race Given Target Race by GPT-4o



The models show some clear preferences for predicting races, with GPT-4 favoring *Asian* and GPT-4o favoring *White* ethnicity, regardless of the content of the essay or the score and rubric provided. The distributions are fairly similar for each *Given Race*, which indicates that both models have certain inherent preferences for race prediction.

For some unknown reason, GPT-4 seems to predict *Asian* much more often than other races, with *White*, *Hispanic*, *Black* following. *Others* was predicted a negligible number of times.

GPT-4o on the other hand, prefers *White* followed by *Asian*, *Black*, *Hispanic* and *Others*. Note that it predicts *Others* in a much higher proportion than GPT-4.

Score Distribution for Predicted Races

Figures 4 and 5 show the predicted race of the generated essays grouped by target score. Recall that the models were given the score and the rubric for the examples essays

before predicting the race and generating the essays. The *Given Score* in the figure refers to that.

Figure 4: Distribution of Predicted Race Grouped by Scores from GPT-4

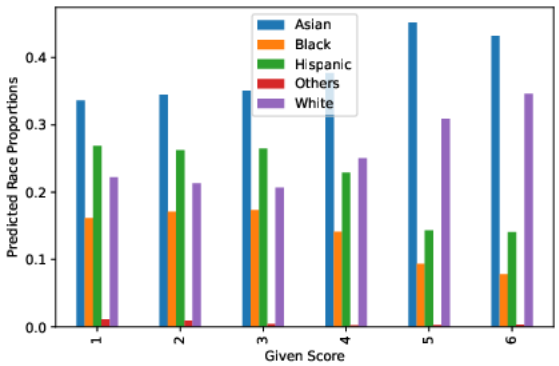
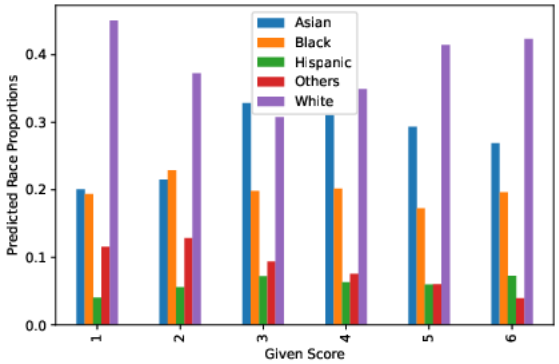


Figure 5: Distribution of Predicted Race Grouped by Scores from GPT-4o



Again as in previous section, for both GPT-4 and GPT-4o, there are clear preferences for race prediction. Both the models predict the race in the same proportions regardless of the score.

Effect of Augmentation on AES

Table 10 contains the QWK obtained by training with and without augmentation for GPT-4 and GPT-4o for both $R = \hat{R}$ and $R \neq \hat{R}$. Except for GPT-4 $R = \hat{R}$, all the models display some appreciation in QWK. The results across subgroups vary but GPT-4o $R \neq \hat{R}$ had the best QWK overall and across 3 of the 6 subgroups.

The results for assessing bias are given in Table 11. We note that the SMD for model trained on original un-augmented data was appreciable across all subgroups, with the base

model (trained on un-augmented data) giving higher mean scores to all subgroups. Augmenting the data decreases the SMD across all subgroups in all cases. GPT-4 $R \neq \hat{R}$ induces a slight negative bias, GPT-4 $R = \hat{R}$ and GPT-4o $R \neq \hat{R}$ have the lowest overall SMD.

Overall GPT-4o $R \neq \hat{R}$ seems to perform well for both increasing the agreement and mitigating bias as measured by QWK and SMD respectively.

Table 10. GPT-4o $R \neq \hat{R}$ as Measured by QWK

	Asian	Black	Hispanic	Others	White	Overall
Non-Augmented	0.806	0.808	0.811	0.808	0.836	0.826
GPT-4 $R = \hat{R}$	0.827	0.799	0.804	0.823	0.838	0.826
GPT-4 $R \neq \hat{R}$	0.805	0.806	**0.814**	**0.839**	0.839	0.829
GPT-4o $R = \hat{R}$	**0.844**	0.816	0.807	0.805	0.828	0.825
GPT-4o $R \neq \hat{R}$	0.834	**0.825**	0.810	0.814	**0.842**	**0.833**

Table 11. GPT-4o $R \neq \hat{R}$ as Measured by SMD

	Asian	Black	Hispanic	Others	White	Overall
Non-Augmented	0.16	0.14	0.16	0.17	0.14	0.15
GPT-4 $R = \hat{R}$	**0.02**	0.03	0.05	0.05	0.03	0.03
GPT-4 $R \neq \hat{R}$	**0.02**	-0.02	**-0.02**	**-0.01**	**-0.01**	**-0.01**
GPT-4o $R = \hat{R}$	0.08	0.09	**0.02**	0.04	0.08	0.06
GPT-4o $R \neq \hat{R}$	0.05	**0.00**	0.02	0.06	0.05	0.03

Discussion and Future Work

We have discussed here, a preliminary study into targeted generation of student essays for AES for mitigating subgroup bias in AES. Our findings reveal that LLMs struggle to internalize the real-world demographic distribution when generating synthetic essays, but the data still improves model performance and helps in bias mitigation. We believe that LLMs have a role to play in AES, but greater controllability in generation and lack of explainability pose significant hurdles. For instance, why exactly has the QWK improved and SMD decreased? That is, which properties of the generated text caused these, needs to be further investigated.

There are several related avenues of research that are possible for work in the future. Better alignment of generated text with human interpretable features would be massively advantageous to understand and trust such data. Future work can also focus on examining which textual cues in given examples to LLM, or in the prompt lead to a particular generated

text for a particular subgroup of interest. We have also left for future work, the study of bias which may be induced in the generated data, and how it may affect the AES system. In addition, we developed our prompt iteratively and did not perform a longitudinal study over its design and constituents, altering which may yield different outcomes.

References

Clark, K., Luong, M.-T., Le, Q. V., & Manning, C. D. (2020, April 26–May 1). *ELECTRA: Pre-training text encoders as discriminators rather than generators* [Paper presentation]. International Conference on Learning Representations 2020, Virtual. https://doi.org/10.48550/arXiv.2003.10555

Crossley, S. A., Tian Y., Baffour P., Franklin A., Benner, M., & Boser, U. (2024). A large-scale corpus for assessing written argumentation: PERSUADE 2.0. *Assessing Writing*, *61*. https://doi.org/10.1016/j.asw.2024.100865

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Long and Short Papers), 1,* 4191–4186. https://doi.org/10.18653/v1/N19-1423

Ding, B., Qin, C., Zhao, R., Luo, T., Li, X., Chen, G., Xia, W., Hu, J., Luu, A., & Joty, S. R. (2024). Data augmentation using LLMs: Data perspectives, learning paradigms and challenges. *Findings of the Association for Computational Linguistics: ACL 2024*, *1*, 1679–1705. https://doi.org/10.18653/v1/2024.findings-acl.97

Fang, L., Lee, G.-G., & Zhai, X. (2023). *Using GPT-4 to augment unbalanced data for automatic scoring.* https://doi.org/10.48550/arXiv.2310.18365

Haberman, S. (2019). *Measures of agreement versus measures of prediction accuracy* (Research Report No. RR-19-20). ETS. https://doi.org/10.1002/ETS2.12258

He, P., Gao, J., & Chen, W. (2023, May 1–May 5). *DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing* [Paper presentation]. International Conference on Learning Representations 2023, Kigali, Rwanda, Africa. https://arxiv.org/pdf/2111.09543.pdf

Johnson, M. S., & McCaffrey, D. F. (2023). Evaluating fairness of automated scoring in educational measurement. In V. Yavena & M. von Davier (Eds.), *Advancing natural language processing in educational assessment* (142–164). Routledge. https://doi.org/10.4324/9781003278658-12

Long, L., Wang, R., Xiao R., Zhao, J., Ding, X., Chen, G., & Wang, H. (2024). On LLMs-Driven synthetic data generation, curation, and evaluation: A survey. In L.-W. Ku, A. Martins, & V. Srikumar (Eds.), *Findings of the Association for Computational Linguistics: ACL 2024* (pp. 11065–11082). Association for Computational Linguistics. https://doi.org/10.48550/arXiv.2406.15126

Loshchilov, I., & Hutter, F. (2017, April 24–26). *Decoupled weight decay regularization* [Paper presentation]. International Conference on Learning Representations 2017, Toulon, France. https://arxiv.org/abs/1711.05101

Morris, W., Holmes, L., Choi, J. S., & Crossley, S. (2024). Automated scoring of constructed response items in math assessment using large language models. *International Journal of Artificial Intelligence in Education, 35*, 559–586. https://doi.org/10.1007/s40593-024-00418-w

OpenAI. (2023). *GPT-4 technical report*. https://arxiv.org/abs/2303.08774

OpenAI. (2024). *GPT-4o system card*. https://doi.org/10.48550/arXiv.2410.21276

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

Yoo, K. M., Park, D., Kang, J., Lee, S.-W., & Park, W. (2021). GPT3Mix: Leveraging large-scale language models for text augmentation. In M.-R. Moens, X. Huang, L. Specia, & S. W.-T. Yih (Eds.), *Findings of the Association for Computational Linguistics: EMNLP 2021* (pp. 2225–2239), Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.findings-emnlp.192

Suggested citation: Badola, A., Zhang, M., & Li, Chen. (in press). On the representation of racial and ethnic subgroups in AI-generated texts: A case study in automated essay scoring. *ETS Research Report Series.* https://doi.org/10.64634/ac01td58