Language Models and Dialect Differences

Jaclyn Ocumpaugh
University of Pennsylvania
Philadelphia, PA, USA
jlocumpaugh@gmail.com

Xiner Liu
University of Pennsylvania
Philadelphia, PA, USA
xiner@upenn.edu

Andres Felipe Zambrano
University of Pennsylvania
Philadelphia, PA, USA
azamb13@upenn.edu

Abstract

The advancements in automatic language processing being ushered in by Large Language Models suggest enormous potential for better personalization during student learning. However, this potential can be best exploited if we know that LLMs are equally capable of interacting with students who speak or write in a range of different dialects. This case study uses systematically manipulated student essays, previously evaluated by human raters, to examine how ChatGPT responds to and addresses specific dialect differences. Results point to important concerns about the potential biases and limitations of both LLMs and humans when evaluating and providing feedback to students who use minoritized dialects. Addressing these concerns is critical for the field of learning analytics, as it seeks to ensure equity and asset-based approaches to learning analytics.

CCS Concepts

• **Applied computing** → Education; Computer-assisted instruction.

Keywords

equity, automatic writing assessment, African American Language, large language models (LLMs)

ACM Reference Format:
Jaclyn Ocumpaugh, Xiner Liu, and Andres Felipe Zambrano. 2025. Language Models and Dialect Differences. In *LAK25: The 15th International Learning Analytics and Knowledge Conference (LAK 2025), March 03–07, 2025, Dublin, Ireland*. ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3706468.3706496

1 Introduction

The potential for large language models (LLMs) to improve human-center learning analytics [38] has been an important topic within the learning analytics research community since their recent inception. In particular, LLMs seem like a promising tool for more accurately recognize students' prior knowledge, a critical step in the design of asset-based approaches to intelligent tutoring design [47]. LLMs have been tested for their effectiveness in personalized feedback and hints generation [39, 52], automated essay scoring [66], plan generation for adaptive scaffolding in student goal setting [24], and quality evaluation of formative assessments [41].

With the enormous power of neural networks and enormous datasets underlying LLMs comes the potential for algorithmic bias (inherent in these data) that can be difficult to overcome. Early chatbots illustrated how quickly blatant forms of racism and sexism might be learned by these algorithms [63], and efforts since have been underway to address these equity issues (e.g., [4, 42]). However, another area of language bias that has been yet to be fully addressed by computational linguists, including those working with large language models, includes the biases toward various dialects that are prevalent across our society [1, 8] and therefore our training data [13, 46].

Biases toward nonstandard dialects are twofold. They include numerical biases in contributions (i.e., smaller speech communities are likely to have contributed far less data than larger speech communities), which mean that meaningful differences in the dialect of smaller groups could be washed out by that of larger groups. Differences in the availability of linguistic data can result in disparities where languages and dialects from less-represented regions receive insufficient attention in model training [21], but they also include instances where speakers whose language is considered prestigious are not adequately served by emerging language technology [36]. As highlighted by recent research, the focus of fairness evaluations and mitigation efforts has largely been on English, while multilingual and non-English languages have received comparatively little attention [51]. Within English, the focus has also been on national-level differences (e.g., English in the US, England, India, etc.), while regional or ethnic dialects like African American Vernacular English have sometimes been actively excluded from these training processes (see Benjamin's [6] discussion of product development at Apple).

In addition to the numeric biases in the training data, LLMs contain attitudinal biases, such as evaluations of language patterns that are implicit in the data. Researchers have documented biases toward women [56], ethnic minorities [60], and political affiliations [12]. Research shows that language models perpetuate covert biases, including those related to dialects, sometimes producing biases that are more pronounced in these models than in human stereotypes recorded in experimental settings [28]. However, there is still limited research on how these covert biases specifically target dialect differences, particularly those affecting marginalized communities.

Emerging research is beginning to address these attitudinal issues in domains other than education. For example, research has shown that African American Language (AAL) patterns are more likely to trigger so-called toxicity measures [43]—incorrectly labeling AAL features as sexist or racist—which obviously raises concerns related to equal access to technology across domains. More recently, Deas et al. [16] examined the how GPT manages AAL grammatical patterns in the context of a chatbot designed to assist

LAK 2025, March 03–07, 2025, Dublin, Ireland

ACM ISBN 979-8-4007-0701-8/25/03
https://doi.org/10.1145/3706468.3706496

in mental health services. However, less work has investigated how LLMs handle AAL in educational contexts.

This paper seeks to engage this problem with a case study that analyzes data from one of the few highly-controlled studies of the perception of AAL in student essays [19]. Specifically, we use Flynn and Preston's [19] essays, which had been manipulated on two dimensions thought to affect student's grades: (a) lexical and syntactic sophistication and (b) the use of AAL. In that study, Flynn and Preston began with an essay that had been published as an exemplar by the New York State's Department of Education (NYS DOE) and labeled it as (1) Good-None (strong lexical/syntactic patterns and no AAL). They then modified to create a (2) Bad-None version, which still had no AAL features but was modified to weaken the essay's lexical and syntactic sophistication. From there each of these essays was modified twice, each time adding more AAL features to generate (3) Good-Low, (4) Bad-Low, (5) Good-High, and (6) Bad-High versions.

This study examines the same six essays to report on an augmented version of Flynn and Preston's [19] data (originally 173 human raters, now 440 human raters), in which each rater saw only one version of the essay. We then examine the evaluations of these essays using GPT-4o (selected because it was the newest model at the time of the analysis) using a zero-shot approach. Our goal is to examine the ability of the LLM to handle the evaluation task in terms of assigning grades, providing general writing feedback, and offering targeted recommendations about grammatical changes (RQ1). We conduct this relatively small-scale study of GPT in the context of a quickly expanding use of LLMs in education, with the goal of better understanding the degree to which algorithmic bias may limit our efforts to improve educational equity (RQ2). We argue that while the number of essays being evaluated by GPT in this study is small, they reflect important human biases that are likely embedded in all LLMs, and we discuss important implications and research questions that emerge when the training data is unlikely to be improved by simply adding more human data.

2 Related Works

Research on student-produced language has typically focused either on essay grading [54] or on chat-bot interactions [10]. Until recently, natural language processing tools required student language to be cleaned to conform to the most prevalent linguistic patterns in the training data [48], often referred to as Standard English. One major advantage of the LLM advances is that they can handle the sorts of minor typos that students might produce with relative ease [62], but less work has investigated the degree to which they manage dialect differences except those that emerge on a national scale (i.e., British vs. American English [53]), with biases towards data from western countries still remaining [32].

In part, this gap in the research is related to limitations in the training data. While it is difficult to know whose language is in the training data of existing LLMs, estimates suggest that African American contributions to common training data sets range from only 0.05 to 6% (see discussion in [16]). Techniques for privileging data from smaller sources exist (i.e., Beltagy's [5] work on finetuning and task adaption strategies for scientific literature and Tran et al.'s [58] work on instruction tuning for medical texts), but—to our knowledge—these have not yet been applied to tackle problems related to equity in non-national dialect differences.

The implications of this bias to the learning analytics community are not well-understood. Although biases toward African American Language more generally are well documented (e.g., [15, 37]), and this has been an important topic in education research at large [17, 57], fewer studies have examined ways in which the designers of intelligent tutoring systems and other online learning systems might accommodate learners from diverse dialect backgrounds. Notable exceptions to this include work to design culturally and linguistically appropriate chatbots [18], which pre-date the advent of LLMs and appear to have been successful in improving student learning and motivation.

Another area of learning analytics research that must consider the implications of dialect differences involves automated essay scoring. Traditional NLP tools have done well at improving student writing, including on measures of cohesion, grammatical structure, and lexical diversity [14, 40], but few studies have looked at the effect of non-standard dialects, like AAL, on essay scoring, and those that have are now quite dated [55]. Work on the use of LLMs to score student essays is just emerging [26, 67]. Warr et al. [59] shows that GPT grades essays more favorably when it assumes that the essay author is White or African American, compared to essays where the information suggests that the author is Hispanic or where that information is assumed to be race neutral. To our knowledge, no previous work has looked at how LLMs might score essays that show marginalized dialect features, but parallel research areas suggest that automated scoring tools may unintentionally disadvantage students who use dialects that diverge from Standard English norms. For instance, research has found that AAL features are often incorrectly labeled as racist or sexist by toxicity filters [43].

Automated essay scoring systems typically assess a range of linguistic and structural features to evaluate the quality of student writing. These features often include grammatical accuracy, structural orderliness, lexical diversity, coherence, and adherence to standard essay organization [50]. For example, such systems might analyze sentence length variation (e.g., [11]), the use of transitional phrases (e.g., [31]), and the sophistication of vocabulary to approximate human grading criteria (e.g., [27]). However, because these systems are generally trained on essays written in Standard English [9], they may not accurately assess writings that incorporate non-standard dialects. This reliance on standard linguistic norms means that dialectal features—such as unique grammatical constructions or region-specific vocabulary—can be misinterpreted as errors or indicators of lower writing proficiency [45]. Consequently, students who use dialects like AAL may receive unfairly low scores due to systemic biases in the assessment algorithms.

This study focuses on how well ChatGPT handles features from AAL as decades of research have shown considerable social biases towards this dialect, which shows consistent grammatical, phonological, and lexical systems that are as rule-governed as national differences that are more widely accepted [44]. Despite its linguistic validity, AAL is still often characterized as a collection of slang or incorrect English rather than a legitimate linguistic variety used by many African American communities [33]. Features of AAL include unique inflections to the verb system that are not found in

standard English. For example, *habitual be* as in "She be working," demonstrates that the woman in question works "all the time" and contrasts with an unmarked form "she working" which denotes that the woman is working "right now." Other AAL features include phonological patterns like consonant cluster reduction, which reduce the number of consonants at the end of the syllable (e.g., "they" for "their" or "col" for "cold"). (See reviews in [25, 44, 64] for more detailed descriptions of this language variety.)

Understanding the ability of LLMs to adequately interact with AAL features is particularly important as research has shown sometimes extreme stigmatization for AAL in educational contexts [3, 29]. Long after researcher produced evidence that US educational system was inadequately serving students who speak AAL (e.g., Labov's [35] classic summary of these concerns), researchers are still trying to ensure that teachers have adequate tools for working with students who use this language variety [2, 22, 61]. Teachers seem to be quite confident in their ability to draw upon students' funds of knowledge, but research suggests that they often have little training on AAL or the linguistic skills necessary to understand its rule-based structure [17]. The consequences can include students whose language is not understood by their teachers, not valued as a resource by their teachers, or (at best) not effectively used as a resource [17, 29, 61]. Under these circumstances, it is important that the learning analytics community ensures that the tools that we are using are not contributing to the problem.

As such, this study investigates the ability of ChatGPT to interact with essays that have been systematically manipulated to introduce AAL features. Specifically, we look at the degree to which AAL features affect the ratings of these essays, but we also examine the degree to which ChatGPT can identify AAL features and distinguish them from grammatical errors. Our goal is to assess whether these tools can accurately interpret student inputs without disadvantaging native AAL speakers, which would perpetuate the sort of educational inequities the learning analytics community hopes to alleviate [30].

3 Methods

This study examines the degree to which features from AAL affect the automated scoring of essays by ChatGPT-4o. To do so, it uses historical data from sociolinguistics research [19], which asked people to score one of six versions of an essay that the New York State Education Department (NYSED) had given an exemplar score in their state standardized tests. Notably, the Flynn & Preston [19] study was presented at a conference that did not issue proceedings, which reduces the risks that GPT might have had prior exposure to these essays during training phases.

3.1 Essay modifications

Flynn and Preston [19] made two types of modifications to the NYSED essay. In the first step, they took the original, good essay and generated a new version that was designed to worsen readability (i.e., generating one additional version of the essay). In the second step, they took both the original and the bad essay, and they modified each to add examples of AAL in varying degrees. This step resulted in three versions of the original essay and three versions of the bad essay, as described in sections 3.1.1 and 3.1.2.

3.1.1 Step 1: Lexical and Syntactic Modifications to Generate the Bad Essay. The first round of modifications took the highly-rated (good) NYSED essay and inserted changes that reduced the syntactic and lexical complexity of the essay without making major changes to its structure or content. These changes followed Freedman's [20] methods for generating sentence-level structures that are predictive of lower essay scores, which are detailed in Table 1. They also reduced the specificity of lexical items, as shown in Table 2. Both the syntactic and the lexical simplification changes are considered emblematic of poorer writing.

The syntactic and lexical modifications introduced in this step did not make major changes to the content of the essay, but they did affect its length. As Table 3 shows, these changes (a) decreased the number of words per sentence in a way that (b) reduced the overall number of words in the essay. By generating shorter, choppier sentences, these changes also (c) increased the number of sentences per paragraph. Other changes weakened the lexical sophistication and specification of the essay (Table 2). At the end of this step, there were two essays: a good essay and a bad essay, neither of which employed any AAL features.

3.1.2 Step 2: Modifications to Insert AAL Features. The second round of modifications were designed to test the effects of AAL on essay scores. In this round, both (1) the original, *Good-None* essay and the (2) newly-minted *Bad-None* essay formed a control condition where no AAL features were present. The essays from the control condition were first modified to include low levels of AAL. This was done by inserting one instance each of eight, well-documented syntactic characteristics of AAL [19], resulting in (3) a *Good-Low* essay and a (4) *Bad-Low* essay. The essays from the control condition were then modified a second time. This time two instances of each AAL feature were inserted, resulting in (5) a *Good-High* essay and (6) a *Bad-High* essay.

3.2 Human Raters

Flynn & Preston [19] reported on judgements from 173 undergraduate students at Michigan State University that were collected in the Spring of 2005. In this study, we have also obtained a parallel data set, collected by Flynn and Preston in 2006, but never reported on in previous literature. This second data set includes judgments from an additional 267 MSU undergrads and has not previously been analyzed. Each student (N=440) was given only one of the six versions of the essay and asked to score the essay as an A, B, C, D, F, without mentioning that these essays might contain AAL features. These scores were then converted to a 4-point grading scale for analysis, where A=4, B=3, C=2, D=1, and F=0.

3.3 Prompt Engineering

Two rounds of prompts were engineered. Prompt engineering is the strategic process of crafting inputs, or prompts, to direct a language model's behavior and responses. This technique has been shown to impact model performance in various ways, including accuracy, relevance, completeness, tone, style, and the level of detail in the output [23]. In our study, we conducted several rounds of prompt modifications to guide the model in producing consistent and reliable results. To mitigate potential biases caused by prior

Table 1: Modifications made to generate a bad essay, based on Freedman [20].

Sentence-level modifications
Include shorter, simpler sentence structures
Include long, rambling sentence structures
Eliminate graceful parallelism
Include verboseness on the sentence level
Use repetitive sentence structures
Eliminate advanced punctuation marks (e.g., semicolon or colon)
Use inappropriate tense
Use inappropriate reference between and within sentences
Eliminate sentence-level modifiers
Misuse words
Include sentence fragments

Table 2: Examples of lexical simplification and reduced specificity

Good Essay	Bad Essay
many advantages and disadvantages	many good and bad things. . .
totalitarian form of government	forms of government with only one person in charge
dictator	Leader
he enforced a totalitarian state	he made all of the rules.
Revolt	Fight
Liberals	some people
totalitarian rulers	rulers that are like kings

Table 3: Basic statistics for the good and bad essays

	Good Essay			Bad Essay		
	Words	**Sentences**	**W/S**	**Words**	**Sentences**	**W/S**
Paragraph 1	56	3	18.7	64	4	16
Paragraph 2	245	12	20.4	241	25	27
Paragraph 3	261	15	15.4	228	27	9.1
Paragraph 4	57	4	14.3	53	5	10.6
Total	619	34	18.2	586	61	9.6

interactions, GPT's memory was cleared at the start of each conversation. Additionally, due to the stochastic nature of GPT's output, we repeated each prompt three times for each analysis to ensure consistency in GPT's responses, but only retain responses from ChatGPT that occur in at least two of those executions. In the first prompt, ChatGPT was (a) given the same information given to the raters in Flynn & Preston [19], which did not include any information about student's ethnicity or dialect background, but ChatGPT was also told that these were 12^{th} grade essays that the students had written under timed conditions where revision was impossible. To ensure that each paper was evaluated separately (as in Flynn & Preston's study, where each rater evaluated only one paper), GPT was explicitly told to treat each essay as a separate submission from different students, as opposed to revisions of the same paper. Afterwards, (b) GPT was asked to identify which grammatical concerns might be attributed to African American Vernacular English (AAVE) patterns in each essay, and (c) to identify grammatical issues that might be attributed to AAVE patterns. This term (AAVE as opposed to AAL) was used in the prompts given to GPT as AAVE has a much longer history of usage in the sociolinguistics literature, even though AAL is becoming the preferred term.

In the second prompt, GPT was given the same initial instructions to evaluate the essays as if they were individual submissions, but it was provided with the list of AAL features it had identified in the third stage of the first prompt, which included (1) double negation, (2) omission of copula, (3) differences in subject/verb agreement, (4) the use of "they" for "their," (5) the simplification of past tense, and the (6) the use of "was" with plural subjects. These features were selected both because they were accurate features of the dialect and because we had established the GPT could identify them. GPT was instructed to grade as if these students were AAL speakers who had been told to write in their own voice. After GPT supplied grades, it was then asked to identify specific grammatical concerns for each essay.

Table 4: Distribution of AAL Characteristics across Essays

AAL Characteristics			Good-None	Bad-None	Good-Low	Bad-Low	Good-High	Bad-High
Existential 'it'	AAL	It's pros and cons to both systems.	0	0	1	1	2	2
	StdEng	There are pros and cons to both systems.						
Double negation	AAL	The citizens don't have no say...	0	0	1	1	2	2
	StdEng	The citizens have no say...						
Plural –s deletion	AAL	It was a responsibility of all AthenianØ to...	0	0	1	1	2	2
	StdEng	It was a responsibility of all Athenians to...						
Possessive –s deletion	AAL	The peopleØ rights are denied.	0	0	1	1	2	2
	StdEng	The people's rights are denied.						
Possessive 'they'	AAL	Both have had they bright points.	0	0	1	1	2	2
	StdEng	Both have had their bright points.						
3rd-person singular –s deletion	AAL	The government representØ the people.	0	0	1	1	2	2
	StdEng	The government represents the people.						
Copula deletion	AAL	He cannot be replaced unless he Ø overthrown.	0	0	1	1	2	2
	StdEng	He cannot be replaced unless he is overthrown.						
Past tense deletion	AAL	This limited the powers of the king and allowØ parliament...	0	0	1	1	2	2
	StdEng	This limited the powers of the king and allowed parliament...						
		Total AAL Characteristics	0	0	8	8	16	16

Table 5: Prompts used in this Study

Stage	Task	Text of Prompt
Prompt 1a	Provide grades (no AAL instructions)	Below are six essays. Please treat each one individually. Do not treat these as if they are revisions of one another, even if they seem very similar. Instead, assume they were written by a 12th grade student who was taking a two-hour standardized exam, rather than a long research paper. Please assign letter grades to each.
Prompt 1b	Provide grammatical feedback	Please provide any additional grammatical concerns you might have for each of the 6 essays. Provide specific examples for any concerns you raise.
Prompt 1c	Identify AAL features	Of the grammatical problems you've identified in these essays, which ones could be attributed to African American Vernacular English patterns?
Prompt 2a	Provide grades (with AAL instructions)	Below are six essays. Please treat each one individually. Do not treat these as if they are revisions of one another, even if they seem very similar. Instead, assume they were written by a 12th grade student who was taking a two-hour standardized exam, rather than a long research paper. Please assign letter grades to each. When you assign these grades, please consider the following specific details about African American Vernacular English (AAVE). AAVE may include (1) double negation, (2) omission of copula, (3) differences in subject/verb agreement, (4) the use of "they" for "their," (5) the simplification of past tense, and the (6) the use of "was" with plural subjects. Assume that the students who have submitted these essays are AAVE speakers who have been told to write in their own voice, and assign the grades accordingly.
Prompt 2b	Provide grammatical feedback	Now please identify any grammatical concerns you might have in each essay. Please provide specific examples of each.

4 results

This section examines the scoring, feedback, and grammatical errors identified by ChatGPT to investigate how different levels of AAL influence the assessment of essay.

4.1 Human Results

As Table 6 shows, the human raters of these essays were generally unimpressed across all six versions. Even though the original essay with no AAL features (Good-None) had been ranked by the NYS DOE as an exemplar, it averaged only a mid C range (2.6) from the human raters. The average ratings went down from there,

Table 6: Descriptive Statistics for Human Ratings of the Six Essays (2005-2006)

	N	Avg	Med	Min	Max	SD	A	B	C	D	F
Good-None	63	2.6	3	1	4	0.92	11	24	20	8	0
Bad-None	79	2.39	2	0	4	0.87	8	26	35	9	1
Good-Low	64	2.04	2	0	4	0.86	2	16	32	11	3
Bad-Low	84	2.02	2	0	4	1.02	56	22	30	20	6
Good-High	77	1.55	2	0	4	0.68	1	3	35	36	2
Bad-High	73	1.82	2	0	4	0.82	2	11	34	24	2

Table 7: Differences for Human Ratings of the Six Essays (2005-2006)

Comparison	P-value	Rank	Adjusted Alpha	Cohen's D
Bad-None vs Bad-Low	0.007*	4	0.011	0.367
Bad-None vs Bad-High	<0.001*	2	0.006	0.675
Good-None vs Good-Low	<0.001*	3	0.008	0.622
Good-None vs Good-High	<0.001*	1	0.003	1.30
Bad-None vs Good-None	0.155	6	0.017	0.646
Bad-Low vs Good-Low	0.874	7	0.019	0.02
Bad-High vs Good-High	0.05	5	0.01	0.27

[a] Asterisks indicate results that were significant after the Benjamini & Yekutieli correction.

driven more strongly by the amount of AAL than by the quality of the essay. The Bad-None essay was only penalized slightly for its reduced syntactic and lexical quality (2.39). The essays with low AAL were penalized more severely (2.04 for the Good-Low essay vs. 2.02 for the Bad-Low Essay). Higher rates of AAL were more stiffly penalized, though this effect was stronger for the essay with strong syntactic and lexical sophistication (1.55 for the Good-High essay vs. 1.82 for the Bad-Essay).

A one-way ANOVA revealed significant differences in the gradings assigned by human raters across essays ($F(5,434)=13.56$, $p<0.001$). Pairwise t-tests showed significant differences between essays of the same quality level that differed only in the inclusion of AAL terms (see Table 7). The Benjamini-Yekutieli correction [7] was applied to control for the family-wise error rate inherent in multiple comparisons. For essays rated as low quality, the inclusion of a low number of AAL terms led to a moderate significant difference in human ratings (Cohen's $d=0.367$, $p=0.007$). A high inclusion of AAL terms resulted in a more substantial difference, with a larger effect size (Cohen's $d=0.675$, $p<0.001$). This difference was even more pronounced for high quality essays. Here, a low inclusion of AAL terms resulted in a moderate-large effect size (Cohen's $d=0.622$, $p<0.001$), while a high inclusion of AAL terms led to a difference greater than one standard deviation between the ratings of the two essays (Cohen's $d=1.30$, $p< 0.001$). Notably, in contrast, no significant differences were observed in the ratings of the two essays that differed in quality but had the same level of inclusion of AAL terms.

4.2 GPT Results

In this section, we report two types of results. First, we report how GPT scored each of the six essays used in Flynn & Preston [19], which we compare to the human ratings collected by those two researchers. Next, we conduct a qualitative examination of the errors identified by GPT, including some which were provided spontaneously, to examine the degree to which AAL features can be identified by GPT. The results reported here are from a single conversation that reflected the overall patterns observed across multiple interactions with GPT.

4.2.1 Scores for each essay, across both prompts. Results from the two prompts show that GPT generally scores these essays better than the human raters did in the 2005-2006 data. GPT ranked "good" versions of the essay higher than "bad" versions of the essay (see Table 8). There is also a downward trend in ratings that is related to the degree to which AAL features have been integrated into essays.

Table 8 also compares the GPT scores to the average scores that each essay received from the human raters. Results indicate that GPT was more consistent than humans when it came to the effects of the modifications related to syntactic and lexical complexity (i.e., the bad essays' modifications). The good-bad differences (Good-Bad_{diff}) ranged from 0.21 to -0.28 for humans, while being 0.67 for the first GPT prompt and 0.33 for the second GPT prompt. In part, these results appear to be driven by human reactions to AAL levels, where human raters graded the good-high essay more harshly than the bad-high essay. In contrast, GPT appears to be more predictable in its reactions to AAL. Prompt 1 ultimately produces larger differences in the range of scores across the six essays (i.e., Prompt 1 scores: 2-3.33; Prompt 2 scores: 2.67-3.67 vs. human scores: 1.55-2.60).

Table 9 shows the impact of AAL levels more directly. Among human raters, the change from no to low levels of AAL has larger effects on the Good essay than on the Bad essay (None-Low_{diff} = 0.56 vs. 0.42), and those differences are consistent when the AAL level increases to high (None-$High_{diff}$ = 1.06 vs. 0.57). For the first

Table 8: Differences between Good and Bad Essay Ratings

Rater	AAL level	Good	Bad	Good-Bad$_{diff}$
Human	None	2.60	2.39	0.21
	Low	2.05	1.98	0.07
	High	1.55	1.82	-0.28
GPT Prompt 1	None	3.33	2.67	0.67
	Low	3	2.33	0.67
	High	2.67	2	0.67
GPT Prompt 2	None	3.67	3.33	0.33
	Low	3.33	3	0.33
	High	3	2.67	0.33

Table 9: Differences between Levels of AAL in Essays

Rater	Essay Quality	None	Low	None-Low$_{diff}$	High	None-High$_{diff}$	Low-High$_{diff}$
Human	Bad	2.39	1.98	0.42	1.82	0.57	0.15
	Good	2.60	2.05	0.56	1.55	1.06	0.50
Prompt 1	Bad	2.67	2.33	0.34	2	0.67	0.33
	Good	3.33	3	0.33	3	0.33	0.00
Prompt 2	Bad	3.33	3	0.33	2.67	0.66	0.33
	Good	3.67	3.33	0.34	3	0.67	0.33

GPT prompt, the effect is uniform when moving from no to low levels of AAL (None-Low$_{diff}$ = 0.3 for both, with differences due only to rounding effects in the operationalization of numeric grades), but only the Bad essay receives a lower grade when transitioning from low to high levels of AAL. For the second GPT prompt, the effect is uniform for both the Good and the Bad essay, with low levels of AAL dropping the letter grade by a third and high levels dropping it another third.

4.2.2 Analysis of Feedback, Prompt 1. Recall that for prompt 1, GPT was asked (in three stages) to (a) evaluate the essays, (b) provide examples of specific grammatical issues, and (c) to identify grammatical issues that might be attributed to AAL patterns. Here we analyze feedback that was provided by prompt 1 in the first two of these examined stages.

In the first stage of prompt 1, GPT provided strengths and weaknesses alongside the grade it assigned for each essay. These took the form of a bulleted list that start with two strengths for each essay and then presented 2-4 areas for improvement. For each strength and weakness, GPT provided a major category (as summarized in Table 10), followed by an explanation that sometimes included an example (e.g., "Structure and Organization: The essay is well-organized, with clear paragraphs and logical transitions between ideas. The structure allows for a coherent argument."). In general, the data show that there is a small effect on the number of weaknesses (or areas of improvement) suggested by GPT and the essay quality, with good essays receiving slightly fewer suggestions for improvement than bad essays. There is also a small effect on the number of weaknesses identified with respect to the amount of AAL present in each essay, with fewer weaknesses identified in the Good-None essay compared to the Good-Low and Good-High essays, and fewer weaknesses identified in the Bad-None and Bad-Low essays than in the Bad-High Essay.

There are also differences in the typology of these weaknesses. Notably, the Good-None essay is the only one not to receive grammar and syntax feedback unprompted. Meanwhile, the Bad-Low and Bad-High essays are the only two to receive feedback on repetition. This finding aligns with the deliberate manipulation of the sentence structure when generating the bad version of the essay, which resulted in many cases of repeated subjects across the new, shorter sentences. However, it was not a criticism in the Bad-None version of the essay.

When asked to explicitly identify grammatical issues in each essay (prompt 1b), GPT generated a list of 34 grammatical issues across the 6 essays, but became confused when trying to identify which sentence went with which essay label. Therefore, Table 11 presents an analysis of the typology of those errors, rather than the results across essays.

As the table shows, not only did GPT struggle to identify the essay it was extracting the error from, it also struggled to correctly name the grammatical issue it was identifying. For example, two instances that it identified as "problems with articles" were in fact cases of subject/verb agreement (i.e., the sentence "this make them want to fight the unfair government"), while a third was from a sentence that might be classified as missing commas under certain style-book requirements. Several other problems identified by the GPT were also unclear, including the classification of sentence in the Bad-None essay (i.e., "It allows parliament to challenge the king's decisions") which was identified as a problem with pronouns, when there is nothing—including the use of pronouns—in the sentence that is incorrect.

Table 10: Strengths and Weaknesses

	Category	Good-None	Good-Low	Good-High	Bad-None	Bad-Low	Bad-High	Grand Total
Strengths	Clarity:	1			1	1	1	4
	Content:	1	1	1	1	1	1	6
	Organization:		1	1				2
	Strengths Subtotal	2	2	2	2	2	2	12
Areas for	Grammar and Syntax:		1	1	1	1	1	5
Improvement	Repetition:					1	1	2
	Analysis:	1				1	1	3
	Conclusion:		1				1	2
	Critical Thinking:			1	1			2
	Depth/Detail and Depth:	1	1	1	1			3
	AoI Subtotal	2	3	3	3	3	4	18
Grand total		4	5	5	5	5	6	30

Table 11: Specific Grammatical Concerns Identified by GPT.

GPT Identified IssueActual Issue	Articles	Negation	Pronoun	Redundancy	Sentence/ Fragment	Verb	Word Choice	Grand Total
Negation*		5				1		6
Pronoun (r-lessness spelling)*			5					5
Conciseness				4	1			5
Verb*	2					6		8
Pronoun (existential it for there)*					1		2	3
Unclear	1	1	1			2	2	7
Grand Total	3	6	6	4	2	9	4	34

[a] Categories marked with a * indicate grammatical features common in AAL.

4.2.3 Analysis of Feedback, Prompt 2. For prompt 2, GPT was asked to score the six essays independently, but it was also told that the students who wrote these essays were AAVE speakers who had been told to write in their own voice. This time GPT only provided two lines of feedback for each essay, one strength (Table 12) and one area of improvement (Table 13).

The strengths identified for each essay show considerably more overlap in the response to Prompt 2 than they did in the response for Prompt 1. The first sentence of each strength identifies the topic (government types), and the second sentence praises the students' use of AAL, even in the two essays in which none was present. Notably, however, GPT names specific features of AAL in only three of these essays: Bad-Low and Bad-High. For the other essays, it simply praises the presence of AAL as evidence that the student is writing in their own voice.

The areas of improvement identified in the response to Prompt 2 are more succinct than Prompt 1. Interestingly, the essays with no AAL are not given any advice that resembles sentence-level concerns, suggesting that GPT is more sensitive to AAL than to sentences that are weaker in terms of syntax and lexical specification. Instead, the Good-None essay writer is advised to add "more specific examples" and "a more through exploration of the drawbacks of each system." Likewise, the Bad-None essay writer is advised to "explore the implications of each government system in more depth" and to work on strengthening the conclusion. This is also true of the Good-High essay, whose writer is advised to provide "more analysis and fewer generalizations." In contrast, other essays that contain AAL are given advice to (a) improve analyses and (b) improve sentence-level qualities.

5 Discussion

This paper has examined the ratings of essays that have been systematically manipulated both for syntactic and lexical sophistication and for levels of AAL. We first examine the effect of these changes on human raters. We then explore the effect of these differences on GPT ratings of the same essays, including a qualitative comparison of the feedback it gives when GPT is prompted to grade these essays without (Prompt 1) and with (Prompt 2) explicit instructions about students' use of this dialect.

By leveraging materials previously used to elicit human ratings, we are able to make inferences about the relationship between human training data and GPT scoring. The historical data provided by Flynn and Preston [19] demonstrates consistent differences in essays based both on their lexical and syntactic sophistication (Good vs. Bad essays) and based on the number of AAL features present in each essay (None, Low, or High). GPT scores largely mirror those results and are perhaps more consistent at privileging the first dimension (sophistication) over the second (dialect differences).

Table 12: Strengths Identified by GPT using Prompt 2

Essay	Strengths
Good-None	This essay is the most polished, with clear organization, strong transitions, and a well-rounded comparison of government systems. The *use of AAVE is present but subtle*, indicating a sophisticated balance between maintaining the student's voice and academic expectations.
Bad-None	This essay is concise and well-organized, with a clear comparison between the two government systems. The *AAVE features are naturally integrated*, maintaining the student's voice.
Good-Low	The essay is well-organized and presents a balanced view of totalitarianism and democracy, using specific historical examples. The *use of AAVE is present but subtly integrated*, showing a good balance between academic tone and personal voice.
Bad-Low	The essay is clear and straightforward in its comparison of different government systems. The *use of AAVE is consistent and natural*, with double negation and subject-verb agreement reflective of the student's voice.
Good-High	This essay presents a well-organized comparison between totalitarianism and democracy, with relevant historical examples. The *use of AAVE features is present but less pronounced*, which suggests some code-switching while maintaining the student's voice.
Bad-High	The essay effectively outlines the differences between totalitarian and democratic governments. The *use of AAVE features* like double negation ("don't have no say") and subject-verb agreement variations ("They might bring economic growth") *is consistent*, indicating the student's voice.

[a] Italics added to highlight discussion of AAVE

Table 13: Areas of Improvement (AoI) Identified by Prompt 2

Essay	Areas of Improvement
Good-None	The essay could be enhanced by adding more specific examples and a more thorough exploration of the drawbacks of each system.
Bad-None	The essay could be improved by exploring the implications of each government system in more depth. The conclusion could be stronger by summarizing the key points more effectively.
Good-Low	While well-argued, the essay could include more in-depth analysis of the consequences of each government system. Some sentences could be revised for clarity.
Bad-Low	Like Essay 6, this essay could benefit from deeper analysis and stronger transitions between ideas. Some sentences are repetitive and could be combined for a more concise argument.
Good-High	The argument could be strengthened by providing more analysis and fewer generalizations, such as "fewer problems" without explaining what those problems are.
Bad-High	The organization could be tighter, and some ideas are repeated without further development. The essay would benefit from clearer transitions and more nuanced arguments.

[a] Underlines added to highlight advice on sentence-level qualities.

GPT also provides higher scores to these essays more generally. When the grading tasks were repeated three times for each prompt, the results demonstrated a clear pattern of consistency across runs. While all three ratings (humans, prompt 1 and prompt 2) showed largely the same types of differences, the Prompt 1 scores were consistently higher than the human raters. This effect was magnified in prompt 2 results, where GPT was told that the students were AAL speakers who were told to write in their own voice. This result suggests that GPT may be more closely replicating the NYS DOE evaluations than the human raters are, since the Good-None version of the essay was identified by that department as an exemplar. While this may feel like a surprising result, the human raters in the Flynn and Preston data were undergraduates at Michigan State University, some of them potentially with little knowledge about AAL, and therefore not experienced graders for this task. By comparison, GPT has likely trained on large numbers of student essays that are available in various locations across the web. If we only look at the holistic scores, GPT has outperformed novice humans by better estimating the Good-None essay score and giving more weight to lexical/syntactic complexity differences than to the use of a non-standard dialect.

That said, GPT did not perform far better than the human raters in terms of its reaction to AAL more broadly. When left unprompted about AAL (Prompt 1), GPT penalized the presence of AAL in the good essays, but the penalty was smaller than that found among the human raters. However, in the bad essays, it showed stronger grade penalties for the presence of these dialect features than the human raters did. This resulted in GPT showing a larger difference between the Good-None essay and the Bad-High essay than the humans. In fact, any amount of AAL resulted in these essays being

penalized by at least a third of a letter grade, even when GPT was explicitly prompted that the students were AAL speakers who had been instructed to write in their own voice (Prompt 2). In other words, a zero-shot approach is insufficient for getting GPT to override its tendency to classify dialect features as errors, which is an important culturally responsive pedagogical strategy.

Other results reveal that GPT may struggle when giving targeted advice to student writers, regardless of their dialect background. Specifically, our analysis of GPT's feedback suggests that it gets confused when trying to identify specific problems with the essays. The analysis of the types of grammatical issues it identified in Prompt 1 shows that it is unable to identify relevant parts of speech for targeted feedback. In extreme cases, it identifies problems in sentences where the language is completely standard and punctuated according to normal classroom conventions.

Interestingly, GPT struggles to provide this feedback despite the ability to adequately describe common features of AAL. When asked at the end of Prompt 1, GPT was able to list known AAL features and provide examples that were relevant to the data it had just analyzed. However, when this list was used to refine the prompt (Prompt 2), GPT's feedback was unable to distinguish between essays that contained AAL and essays that did not. Instead, it complimented the hypothetical writers of all six essays on showing their voice by using AAL—even as its grades continued to penalize the presence of actual AAL features.

6 Conclusions

To our knowledge, this is the first paper to examine GPT's ability to handle AAL in student essays, and it does so alongside previously unreported results on how human raters would perform when provided with the same stimulus. Our results have shown that GPT's wholistic scoring of essays is potentially useful, but its feedback on strengths and weaknesses likely needs additional fine tuning. These results raise a number of questions about equity and fairness when using GPT specifically and in essay scoring more broadly.

Currently, the LLMs are better able to manage the messy data that children and other users are prone to producing, allowing them to function without the extensive data cleaning that traditional NLP tools have required. This ability could potentially help LLMs to better serve students whose language patterns do not match the classroom standard, but our results suggest that we should proceed cautiously. Although the LLMs are capable of interacting with this data, they are still reproducing human biases against non-standard English, and they are unable to articulate exactly what they are reacting to when they do. It would appear from our efforts to pilot this research that GPT can list the features of AAL, but thus far, its skills at responding in a culturally responsive manner are limited to praising its use (whether or not it was present). We do not yet know whether this is something that would replicate at scale (with a large number of essays) or across other minoritized dialects that have also faced social stigma (e.g., Chicano English).

We also do not know what it will take to ensure that it is capable of appropriately interacting with students from minoritized dialects in a culturally-responsive manner. It is clear that GPT is a very powerful tool, and extensive prompt engineering or fine-tuning could improve its performance in this space. However, it is possible that the amount of AAL in the training data is so low that GPT will need to be explicitly re-tuned on data that matches those patterns, which would have fairness implications that extend well beyond educational contexts. Evidence both from this study and from research showing the propensity of toxicity algorithms to mis-label AAL as racist or misogynist suggest that the latter process may be required.

This small-scale study (six essays) leverages data that was explicitly designed to investigate the degree to which human raters would react more harshly to AAL features than to syntactic and lexical differences that are known to reduce essay quality irrespective of dialect differences [19]. AAL has sometimes been characterized as one of the most stigmatized varieties of English [65], so this work represents one piece in a large body of sociolinguistic research on discrimination against AAL speakers both within education and beyond (e.g.. [3]). Therefore, addressing AAL differences in educational contexts is an important step towards improving educational equity, but ensuring that we have fully equitable access to educational technology will require addressing differences that represent a wide range of regional and ethnic backgrounds. Future research should consider how these efforts could be applied to the language norms of other minoritized groups in the service of the same equity-based goals.

Beyond the ability to recognize and respond to AAL differences, the field should also consider how GPT might change educational goals. If GPT were able to accurately convert AAL features to grammatical patterns that are more widely accepted, then perhaps the use of GPT as a tool for standardization should be encouraged in the same way that word processing applications now check for both spelling and grammar. This kind of use could improve student engagement by de-stigmatizing the translation process between minoritized and standardized language varieties, while also allowing teachers to focus on the development of ideas, which are often why educators assign essays in the first place. Legitimate concerns about the homogenization of student language [34, 49] should be carefully considered, but using these tools to promote students' ability to be linguistically flexible is an important educational goal.

The field is clearly on the cusp of major changes in education that are already emerging as a result of LLMs. As the changes occur, it is important that we ensure that these tools are being used in a way that empowers students and encourages agency. With advances in our understanding of how LLMs process minoritized dialects, the field should be able to support and encourage those uses in the classroom.

Acknowledgments

We posthumously thank and acknowledge Dan Flynn, whose work inspired and enabled this paper. We thank Dennis Preston for sharing the MSU data with us. All opinions are our own.

References

[1] Carolyn Temple Adger, Walt Wolfram, and Donna Christian. 2014. *Dialects in schools and communities*. Routledge.

[2] Jeff Bale, Shakina Rajendram, Katie Brubacher, Mama Adobea Nii Owoo, Jennifer Burton, Yiran Zhang, Elizabeth Jean Larson, Antoinette Gagné, and Julie Kerekes. 2023. *Centering multilingual learners and countering raciolinguistic ideologies in teacher education: Principles, policies and practices*. Channel View Publications.

[3] John Baugh. 1999. *Out of the mouths of slaves: African American language and educational malpractice.* University of Texas Press.

[4] Hedin Beattie, Lanier Watkins, William H. Robinson, Aviel Rubin, and Shari Watkins. 2022. Measuring and Mitigating Bias in AI-Chatbots. In *2022 IEEE International Conference on Assured Autonomy (ICAA)*, 2022. 117–123.

[5] Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676* (2019).

[6] Ruha Benjamin. 2023. Race after technology. In *Social Theory Re-Wired*. Routledge, 405–415.

[7] Yoav Benjamini and Daniel Yekutieli. 2001. The Control of the False Discovery Rate in Multiple Testing under Dependency. *The Annals of Statistics* 29, 4 (2001), 1165–1188.

[8] Camiel J. Beukeboom and Christian Burgers. 2019. How stereotypes are shared through language: a review and introduction of the aocial categories and stereotypes communication (SCSC) framework. *Review of Communication Research* 7, (2019), 1–37.

[9] Jill Burstein and Martin Chodorow. 1999. Automated essay scoring for nonnative English speakers. In *Computer mediated language assessment and evaluation in natural language processing*, 1999.

[10] Fidel Çakmak. 2022. Chatbot-Human Interaction and Its Effects on EFL Students' L2 Speaking Performance and Anxiety. *Novitas-ROYAL (Research on Youth and Language)* 16, 2 (2022), 113–131.

[11] Hongbo Chen and Ben He. 2013. Automated essay scoring by maximizing human-machine agreement. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, 2013. 1741–1752.

[12] Wei-Fan Chen, Henning Wachsmuth, Khalid Al-Khatib, and Benno Stein. 2018. Learning to Flip the Bias of News Headlines. In *Proceedings of the 11th International Conference on Natural Language Generation*, November 2018. Association for Computational Linguistics, Tilburg University, The Netherlands, 79–88.

[13] Emma Louise Clark, Catherine Easton, and Sarah Verdon. 2021. The impact of linguistic bias upon speech-language pathologists' attitudes towards non-standard dialects of English. *Clinical Linguistics & Phonetics* 35, 6 (June 2021), 542–559.

[14] Scott Crossley and Danielle McNamara. 2010. Cohesion, coherence, and expert evaluations of writing proficiency. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, 2010. .

[15] Jamell Dacon, Haochen Liu, and Jiliang Tang. 2022. Evaluating and Mitigating Inherent Linguistic Bias of African American English through Inference. In *Proceedings of the 29th International Conference on Computational Linguistics*, October 2022. International Committee on Computational Linguistics, Gyeongju, Republic of Korea, 1442–1454. Retrieved from https://aclanthology.org/2022.coling-1.124

[16] Nicholas Deas, Jessi Grieser, Shana Kleiner, Desmond Patton, Elsbeth Turcan, and Kathleen McKeown. 2023. Evaluation of African American language bias in natural language generation. *arXiv preprint arXiv:2305.14291* (2023).

[17] Emily A. Diehm and Alison Eisel Hendricks. 2021. Teachers' Content Knowledge and Pedagogical Beliefs Regarding the Use of African American English. *Language, Speech, and Hearing Services in Schools* 52, 1 (January 2021), 100–117. https://doi.org/10.1044/2020_LSHSS-19-00101

[18] Samantha Finkelstein, Evelyn Yarzebinski, Callie Vaughn, Amy Ogan, and Justine Cassell. 2013. The Effects of Culturally Congruent Educational Technologies on Student Achievement. In *Artificial Intelligence in Education*, 2013. Springer Berlin Heidelberg, Berlin, Heidelberg, 493–502.

[19] Danny Flynn and Dennis Preston. 2005. 'It's good things and bad things to both systems': AAE morphosyntactic features and educational Evaluation. In *New Waves of Analyzing Variation (NWAV 33)*, 2005. New York University.

[20] Sarah Freedman. 1978. The Evaluators of Student Writing. (1978).

[21] Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed. 2024. Bias and Fairness in Large Language Models: A Survey. *Computational Linguistics* 50, 3 (September 2024), 1097–1179.

[22] Brandy Gatlin-Nash, Lakeisha Johnson, and Ryan Lee-James. 2020. Linguistic differences and learning to read for nonmainstream dialect speakers. *Perspectives on Language and Literacy* 46, 3 (2020), 28–35.

[23] Louie Giray. 2023. Prompt Engineering with ChatGPT: A Guide for Academic Writers. *Annals of Biomedical Engineering* 51, 12 (December 2023), 2629–2633.

[24] Alex Goslen, Yeo Jin Kim, Jonathan Rowe, and James Lester. 2024. LLM-Based Student Plan Generation for Adaptive Scaffolding in Game-Based Learning Environments. *International Journal of Artificial Intelligence in Education* (July 2024).

[25] Lisa J Green. 2002. *African American English: a linguistic introduction.* Cambridge University Press.

[26] Reto Gubelmann, Michael Burkhard, Rositsa V. Ivanova, Christina Niklaus, Bernhard Bermeitinger, and Siegfried Handschuh. 2024. Exploring the Usefulness of Open and Proprietary LLMs in Argumentative Writing Support. In *Artificial Intelligence in Education. Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners, Doctoral Consortium and Blue Sky*, 2024. Springer Nature Switzerland, Cham, 175–182.

[27] George Higginbotham and Jacqui Reid. 2019. The lexical sophistication of second language learners' academic essays. *Journal of English for Academic Purposes* 37, (January 2019), 127–140.

[28] Valentin Hofmann, Pratyusha Ria Kalluri, Dan Jurafsky, and Sharese King. 2024. Dialect prejudice predicts AI decisions about people's character, employability, and criminality. *arXiv preprint arXiv:2403.00742* (2024).

[29] Sharroky Hollie. 2017. Culturally and linguistically responsive teaching and learning: Classroom practices for student success. Teacher Created Materials.

[30] Kenneth Holstein and Shayan Doroudi. 2022. Equity and artificial intelligence in education. In *The ethics of artificial intelligence in education*. Routledge, 151–173.

[31] Shih-Jen Huang. 2014. Automated versus Human Scoring: A Case Study in an EFL Context. *Electronic Journal of Foreign Language Teaching* 11, (2014).

[32] Rebecca L Johnson, Giada Pistilli, Natalia Menédez-González, Leslye Denisse Dias Duran, Enrico Panai, Julija Kalpokiene, and Donald Jay Bertulfo. 2022. The Ghost in the Machine has an American accent: value conflict in GPT-3. *arXiv preprint arXiv:2203.07785* (2022).

[33] Sharese King. 2020. From African American Vernacular English to African American Language: Rethinking the Study of Race and Language in African Americans' Speech. *Annual Review of Linguistics 6*, 285–300.

[34] Maria Kuteeva and Marta Andersson. 2024. Diversity and Standards in Writing for Publication in the Age of AI—Between a Rock and a Hard Place. *Applied Linguistics* (2024).

[35] William Labov. 1970. The Logic of Nonstandard English. In *Language and Poverty*, Frederick Williams (ed.). Academic Press, 153–189.

[36] Nina Markl. 2022. Language variation and algorithmic bias: understanding algorithmic bias in British English automatic speech recognition. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (*FAccT '22*), 2022. Association for Computing Machinery, New York, NY, USA, 521–534.

[37] Joshua L Martin and Kelly Elizabeth Wright. 2023. Bias in automatic speech recognition: The case of African American language. *Applied Linguistics* 44, 4 (2023), 613–630.

[38] Elisabetta Mazzullo, Okan Bulut, Tarid Wongvorachan, and Bin Tan. 2023. Learning Analytics in the Era of Large Language Models. *Analytics* 2, 4 (2023), 877–898.

[39] Jennifer Meyer, Thorben Jansen, Ronja Schiller, Lucas W. Liebenow, Marlene Steinbach, Andrea Horbach, and Johanna Fleckenstein. 2024. Using LLMs to bring evidence-based feedback into the classroom: AI-generated feedback increases secondary students' text revision, motivation, and positive emotions. *Computers and Education: Artificial Intelligence* 6, (June 2024), 100199.

[40] Atsushi Mizumoto and Masaki Eguchi. 2023. Exploring the potential of using an AI language model for automated essay scoring. *Research Methods in Applied Linguistics* 2, 2 (2023), 100050.

[41] Steven Moore, Huy A Nguyen, Tianying Chen, and John Stamper. 2023. Assessing the quality of multiple-choice questions using gpt-4 and rule-based methods. In *European Conference on Technology Enhanced Learning*, 2023. Springer, 229–245.

[42] Kirsten Morehouse, Weiwei Pan, Juan Manuel Contreras, and Mahzarin R. Banaji. 2024. Bias Transmission in Large Language Models: Evidence from Gender-Occupation Bias in GPT-4. In *ICML 2024 Next Generation of AI Safety Workshop*, 2024.

[43] Marzieh Mozafari, Reza Farahbakhsh, and Noël Crespi. 2020. Hate speech detection and racial bias mitigation in social media based on BERT model. *PLOS ONE* 15, 8 (August 2020), e0237861.

[44] Salikoko S Mufwene, John R Rickford, Guy Bailey, and John Baugh. 2001. *What is African American English?*

[45] Jaylin N Nesbitt. 2022. Writing while Black: African American vernacular English (AAVE) and perceived writing performance. (2022).

[46] Mikel K. Ngueajio and Gloria Washington. 2022. Hey ASR System! Why Aren't You More Inclusive? In *HCI International 2022 – Late Breaking Papers: Interacting with eXtended Reality and Artificial Intelligence*, 2022. Springer Nature Switzerland, 421–440.

[47] Jaclyn Ocumpaugh, Rod D. Roscoe, Ryan S. Baker, Stephen Hutt, and Stephen J. Aguilar. 2024. Toward Asset-based Instruction and Assessment in Artificial Intelligence in Education. *International Journal of Artificial Intelligence in Education* (January 2024).

[48] Tor Ole B. Odden, Alessandro Marin, and John L. Rudolph. 2021. How has Science Education changed over the last 100 years? An analysis using natural language processing. *Science Education* 105, 4 (July 2021), 653–680.

[49] Ameena L Payne, Tasha Austin, and Aris M Clemons. 2024. Beyond the front yard: The dehumanizing message of accent-altering technology. *Applied Linguistics* (2024).

[50] Dadi Ramesh and Suresh Kumar Sanampudi. 2022. An automated essay scoring systems: a systematic literature review. *Artificial Intelligence Review* 55, 3 (March 2022), 2495–2527.

[51] Krithika Ramesh, Sunayana Sitaram, and Monojit Choudhury. 2023. Fairness in language models beyond English: Gaps and challenges. *arXiv preprint arXiv: 2302.12578* (2023).

[52] Lianne Roest, Hieke Keuning, and Johan Jeuring. 2024. Next-Step Hint Generation for Introductory Programming Using Large Language Models. In *Proceedings of the 26th Australasian Computing Education Conference* (*ACE '24*), 2024. Association for Computing Machinery, New York, NY, USA, 144–153.

[53] Günter Rohdenburg and Julia Schlüter. 2009. *One language, two grammars?: Differences between British and American English*. Cambridge University Press.
[54] Amit Rokade, Bhushan Patil, Sana Rajani, Surabhi Revandkar, and Rajashree Shedge. 2018. Automated Grading System Using Natural Language Processing. In *2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT)*, 2018. 1123–1127.
[55] MD Shermis. 2003. Automated essay scoring: A cross-disciplinary perspective. *Inc., Publishers. Mahawah* (2003).
[56] Rachael Tatman and Conner Kasten. 2017. Effects of Talker Dialect, Gender & Race on Accuracy of Bing Speech and YouTube Automatic Captions. In *Interspeech*, 2017. 934–938.
[57] Orlando Taylor and Dorian Latham Lee. 1987. Standardized tests and African Americans: Communication and language issues. *The Negro Educational Review* 38, 2 (1987), 67.
[58] Hieu Tran, Zhichao Yang, Zonghai Yao, and Hong Yu. 2024. BioInstruct: instruction tuning of large language models for biomedical natural language processing. *Journal of the American Medical Informatics Association* (2024).
[59] Melissa Warr, Nicole Jakubczyk Oster, and Roger Isaac. Implicit bias in large language models: Experimental proof and implications for education. *Journal of Research on Technology in Education*, 1–24.
[60] Alicia Beckford Wassink, Cady Gansen, and Isabel Bartholomew. 2022. Uneven success: automatic speech recognition and ethnicity-related dialects. *Speech Communication* 140, (May 2022), 50–70.
[61] Rebecca Wheeler. 2016. "So Much Research, So Little Change": Teaching Standard English in African American Classrooms. *Annual Review of Linguistics 2*, 367–390.
[62] Edward Whittaker and Ikuo Kitagishi. 2024. Large Language Models for Simultaneous Named Entity Extraction and Spelling Correction. *arXiv preprint arXiv:2403.00528* (2024).
[63] M. J. Wolf, K. Miller, and F. S. Grodzinsky. 2017. Why we should have seen that coming: comments on Microsoft's tay "experiment," and wider implications. *SIGCAS Comput. Soc.* 47, 3 (September 2017), 54–64.
[64] Walt Wolfram and Natalie Schilling. 2015. *American English: dialects and variation*. John Wiley & Sons.
[65] Walt Wolfram and Erick Thomas. 2008. *The Development of African American English*. John Wiley & Sons.
[66] Changrong Xiao, Wenxing Ma, Sean Xin Xu, Kunpeng Zhang, Yufang Wang, and Qi Fu. 2024. From Automation to Augmentation: Large Language Models Elevating Essay Scoring Landscape. *arXiv preprint arXiv:2401.06431* (2024).
[67] Fatih Yavuz, Özgür Çelik, and Gamze Yavaş Çelik. 2024. Utilizing large language models for EFL essay grading: An examination of reliability and validity in rubric-based assessments. *British Journal of Educational Technology* n/a, n/a (June 2024).