Nonstandard English and the Automated Scoring of Open-Ended Math Problems

Abubakir Siedahmed
Worcester Polytechnic Institute
asiedahmed@wpi.edu

Jaclyn Ocumpaugh
University of Pennsylvania
ojaclyn@upenn.edu

Zelda Ferris
Worcester Polytechnic Institute
zferris@wpi.edu

Dinesh Kodwani
Worcester Polytechnic Institute
dkodwani@wpi.edu

Eamon Worden
Worcester Polytechnic Insititue
elworden@wpi.edu

Neil Heffernan
Worcester Polytechnic Institute
nth@wpi.edu

ABSTRACT

Recent advances in AI have opened the door for the automated scoring of open-ended math problems, which were previously much more difficult to assess at scale. However, we know that biases still remain in some of these algorithms. For example, recent research on the automated scoring of student essays has shown that certain varieties of English are more strongly penalized for non-standard English than they are for other differences that reduce the quality of students' writing. This study examines that issue in a new domain, investigating the potential for large language models to accurately grade open-ended math problems produced by students who speak and write in non-standard English. Specifically, we look at four features of African American Vernacular English (AAVE), which range in the degree to which they are unique to AAVE or are common in other non-standard dialects. We then compare the scoring of answers that were produced by students using these dialect features to a control group of synthetic data--where we converted all non-standard dialect features to standard English. Results show that minor changes in the number of dialect features per student response do not impact GPTs automated scoring, but prompt engineering efforts did.

Keywords

Automated math scoring, large language models, non-standard dialects, fair scoring models, automated content scoring.

1. INTRODUCTION

The rise of Large Language Models (LLMs) has opened new avenues for educators to improve student learning. Researchers have explored LLMs' ability to provide immediate and personalized feedback [12, 31], act as a tutor [45], and generate culturally sensitive and motivational feedback messages [3]. Beyond supporting students, LLMs can also aid teachers by automating tasks, such as grading and feedback generation. Teachers spend significant time grading essays and providing students with feedback [41]. LLMs offer a potential way to automate grading and feedback in ways that could provide students with timely pedagogical support, but our abilities to do that effectively will be hampered if biases in these models keep them from recognizing the language patterns of students from non-standard dialects.

Much of the work on student language recognition has been within the space of automated essay scoring (AES). Until recently, many of these studies used a combination of machine learning and other artificial intelligence (AI) techniques [44]. Since the advent of LLMs, this work has expanded to test their potential for AES [32, 55], but little research has done so within the context of non-standard dialect features, such as those associated with African American Vernacular English (AAVE).

Long before generative AI, AAVE drew interest from those who were concerned about how its treatment was contributing to educational disparities [3, 9, 10, 13, 28, 29, 30, 40, 51] In fact, AAVE is the most studied English dialect [42]. A number of well-documented grammatical features of AAVE, such as the *habitual b*e and *preterit done*, are unique and differentiate themselves from Standard American English (SAE) varieties [15, 34]. Despite these differences, it is vital to indicate that AAVE is not broken or improper English [6, 11, 29, 54]. Instead, it is a legitimate dialect that originated from historical and cultural circumstances [34, 42, 52, 53]. There have been serious efforts to integrate this information into pedagogical training and the educational research community [6, 21], with concerns ranging from the degree to which teachers are able to understand distinct patterns well enough to assess language-related competences [28] to more general worries about stigmatization processes.

In recognizing the importance of acknowledging AAVE features in educational contexts, recent work has explored natural language processing techniques to identify and analyze AAVE features in student writing. For example, studies have shown the presence of AAVE features does not correlate with lower writing quality, but there was a negative relationship with writing performance [35]. More recent work systematically manipulated essays to include and exclude AAVE features to examine how LLMs, like ChatGPT, respond to dialect differences, demonstrating scoring biases [36].

To the best of our knowledge, this study is the first to investigate LLMs' ability to score open-ended mathematical responses containing AAVE features. Using data from a computer-based learning platform, we evaluate GPT-4o performance in scoring responses with AAVE features compared to the same response but using SEA instead of AAVE. We used a zero-shot prompt approach since many LLM-based automated feedback systems used in authentic

Abubakir Siedahmed, Jaclyn Ocumpaugh, Zelda Ferris, Dinesh Kodwani, Neil Heffernan, and Eamon Worden. Nonstandard English and the Automated Scoring of Open-Ended Math Problems. In Caitlin Mills, Giora Alexandron, Davide Taibi, Giosuè Lo Bosco, and Luc Paquette (eds.) Proceedings of the 18th International Conference on Educational Data Mining, Palermo, Italy, July, 2025, pp. 254–264. International Educational Data Mining Society (2025).

https://doi.org/10.5281/zenodo.15870175

learning environments will not adjust their prompt to account for AAVE.

In this study, we aim to answer the following research questions:

RQ1: What temperature setting should be used for GPT-4o to ensure reliable scoring of mathematical open-ended responses containing AAVE and SAE features?

RQ2: How does GPT-4o score mathematical responses with AAVE features compared to the same response with SAE?

RQ3: Does explicitly indicating the presence of AAVE features in the prompt affect GPT-4o scoring of responses containing AAVE?

In doing so, we hope to contribute to the research on how LLMs might be used to develop effective, targeted, and personalized learning instruction.

2. LITERATURE REVIEW

In the short time that LLMs have been available, researchers have already begun to explore their potential for education [33]. Although there have been some ethical concerns about the use of LLMs in education, including their potential to increase the homogenization of language in online learning systems, much of the research has been promising.

To date, the focus in the EDM community has been on the applications [7, 46] rather than the underlying structures and what those implications might be. The neural nets underlying GPT's structure are generally black boxes, and interpreting their mathematical structure is not a plausible task. However, we do know that these models rely heavily on dependencies--including common relationships across words and sentences--to make their predictions [48]. This means that LLMs are able to make some predictions even when there are unexpected grammatical patterns, but it may also mean that if they have been heavily trained on a dialect that uses one set of agreement patterns (e.g., subject verb agreement) they could be more susceptible to errors when evaluating a dialect that diverges from those patterns

That is, what we gain from training large data sets primarily on one dialect pattern may cause the system to override less common patterns. This problem may be particularly pernicious for verbal inflectional patterns, for example, which are relatively ubiquitous in terms of their distribution pattern in a dialect. If the LLM is relying upon dependencies that are less common in the training data, it may have more errors when presented with a dialect different to the training data.

The underlying models for GPT have been adapted for some dialects (American vs. British English vs Indian, for example). The processes involved in this adaptation are proprietary, but likely involve data augmentation and fine tuning, though other kinds of retraining may also be possible. In many ways, they may also be similar to the processes required to train LLMs to learn the jargon and specific knowledge required for these models to be useful in extracting medical and legal data.

Still, we know that it has not been adapted for every possible unique variety of English. In fact, researchers who have worked in industry raise concerns about the lack of training for African American Vernacular English specifically because some who were in positions to set priorities about this kind of adoption did not think that speakers of this variety constituted an important customer base [5]. Because they likely have not been explicitly trained on AAVE patterns, the education community should proceed cautiously when applying them to the data produced by AAVE speakers.

Although many grammatical features of AAVE are not completely unique to that dialect--which has maintained contact with many varieties of American English throughout its history--some features are known to be uninterpretable to people who did not grow up speaking the dialect. In other words, even within the language produced by native speakers of AAVE, some features could be better recognized by LLMs than others. (See Table 1).

Table 1. Four Grammatical Features of AAVE

Construct	Definition and Examples
Negative concord	**Def.** Agreement across the sentence so that negation of the verb is also included in the associated noun phrases. Sometimes referred to as *double negation*, but as Example A [15] shows, it may involve more than two instances of negation. **Example A:** I don't never have no problems. **Gloss:** 'I don't ever have any problems' **Example B:** He didn't do no homework. **Gloss:** 'He didn't do homework.
S/V agreement leveling	**Def.** The lack of marking that shows agreement between a subject and verb in standard English. This can be seen in both past and present forms of the copula in AAVE (where *was* and *is* are used for all subjects) and in 3rd person present (where -s would be attached to verbs in Standard English but is absent in AAVE). **Example A:** They is not the same. **Gloss:** 'They are not the same.' **Example B:** The problems was hard. **Gloss:** 'The problems were hard.' **Example C:** She want harder problems. **Gloss:** 'She wants harder problems.'
Preterit done	**Def.** The use of *done* for *did* in front of nouns to indicate past tense or completed action. **Example A:** I done it this way. **Gloss:** 'I did it this way' **Example B:** They done that problem set already. **Gloss:** 'They did that problem set already.'
Habitual be	**Def.** The use of unconjugated *be* to indicate continuous, habitual, or repetitive pattern that could be compatible with adverbial phrases like "sometimes" or "always." **Example A:** They be on the TV all night. **Gloss:** 'They are on the TV all night' **Example B:** They be similar **Gloss:** 'They are similar' **Example C:** Sometimes they be $20. **Gloss:** 'Sometimes they are $20.'

Examples of features that might be well-recognized by LLMs include negative concord (colloquially referred to as double negatives) where negation is marked in multiple places across the sentences to show agreement, compared to in the current preferences of mainstream American English, where it is only included in the verb phrase. This feature is common cross-linguistically (meaning that LLMs have certainly encountered it in non-English

languages), but also in many varieties of American English, including those spoken by many White Americans in the South. As a consequence, it is likely used and understood by many people who do not speak AAVE.

Likewise, both subject/verb agreement and preterit done may be present in other English dialects [8, 17, 19, 26, 43, 47, 50]. Subject verb agreement is common cross-linguistically, as it produces more opportunities in the speech signal for important content to be perceived. Still, this kind of agreement is not language universal. In fact, even standard English shows relatively weak patterns of agreement compared to other languages and its own historical origins. Thus, this kind of leveling (loss of subject/verb agreement) is likely common in LLM training data, even though the specific paradigm of done for did may be more common in AAVE than other non-standard varieties.

In contrast, AAVE also exhibits what is called a split copula system, which readers may have encountered in other languages (e.g., Spanish, Irish, Hebrew, Japanese, Yorobu, and others), but which is not common in other varieties of American English [14, 34, 52]. That is, AAVE has multiple ways of using the verb *to be*—otherwise known as a copula. Specifically, there is a distinction between what is called the null copula, where the verb is not pronounced, and what is called habitual be, where the unconjugated form is used to indicate habitual aspect—a paradigm that often requires adverbial construction in standard English. As a result, speakers of AAVE recognize that the sentence "She busy" means that "She is busy right now" while "She be busy" means "She is busy all the time" — a distinction that non-AAVE speakers are unaware of [25]. Moreover, this characteristic tends to be relatively rare [39]. As such, if an LLM were not trained explicitly on this feature, it might be more likely to treat it as a typo.

To date, we are not aware of any research that has looked at the degree to which these grammatical differences might affect the grading of math problems, but emerging research has begun to look at the effect that they might have on the automated grading of student essays. Building on earlier work that systematically manipulated an essay to test human biases (and showed that AAVE grammatical features were more steeply penalized than bad essays), [36, 37] have begun to explore this issue using GPT-4.

In this work, researchers found that the chatbot was slightly less biased than humans on scoring but struggled to provide targeted feedback in part because it was misidentifying parts of speech [36]. A combination of prompt engineering [36, 37] and adjustments to the temperature parameters in GPT's API [37] improved this performance. Namely, lower temperature parameters resulted in consistently lower scores for bad essays than for the presence of AAVE in good essays [37], but lower temperatures did not improve the LLM's ability to identify parts of speech--a key component of the feedback it was inclined to provide. Instead, feedback improved at higher temperature parameters, but was still biased towards sentences that contained AAVE features (Ocu-ST&D). These findings were ameliorated by prompt engineering that specifically described AAVE features and asked the LLM to treat these essays as if the students were told to write in their own voice [36, 37], but they were not eliminated.

Although tested in a humanities learning domain, these findings have important implications for writing in STEM domains where there are sometimes higher expectations about conformity to standard language patterns than you might see in the humanities (e.g., where poetry and other genres sometimes encourage more creative uses of language). If LLMs are more highly attuned to presence of AAVE than to issues related to poor writing (weaker lexical and syntactic choices), these biases are likely domain-independent. Therefore, it is important to test how these might emerge in domains like mathematics, where training data from experts is even less likely to include creative and other non-conformist language patterns.

3. METHODS

3.1 ASSISTments

The data used for this study is from a computer-based learning platform (CBLP) called ASSISTments [18]. ASSISTments is one of the most widely used CBLPs in the United States, with over one million annual student users and over 30,000 teachers [2]. The platform allows teachers to assign middle and high school mathematical problems to students. Teachers can select which problems they want to assign from various in-house open-source curriculums, such as Illustrative Mathematics or EngageNY. Teachers can also create their own problems. Students are required to complete each problem before moving on to the next problem. If a student is stuck on a problem, they're able to request for assistance, which would appear in the form of a hint or an explanation.

Problems in ASSISTments can appear in the form of several types. There are drag and drop problem types, where students have to select an answer from a set of options and drag the answer and drop it in the solution text box. Other problem types are fill in the blank, multiple choice, and open-responses questions. Most problem types are computer-gradable, with the exception of open-response questions. Most open-response questions in ASSISTments are conceptual, which means that a student is prompted to *explain their reasoning* or *explain why or why not*. Developing automated graders for this particular problem type is an important goal because of the potential it has to improve learning. To date, these problems must still be hand-scored by teachers, which can slow the time it takes for struggling students to receive feedback.

Typically, conceptual problems are found in every assigned problem set to students. These are hand-scored, the grading scale may be different from one set of teachers to the next, but some teachers follow grading scales set by their district or curriculum. One such rubric is that of Illustrative Mathematics, which is presented in greater detail below (see Section 3.4). It uses a five-point rubric that ranges from 0-4 (poor to strong answers), which would result in a maximum of 20 points if there were 5 problems, and 100 points if there were 25 problems. We highlight this point to remind readers that this means that a reduction of scores on individual problems for math-irrelevant issues (e.g., grammatical differences) could accumulate quickly, such that even a 0.2 discrepancy could lower a child's assignment-level score by half a letter grade under some calculation schema (e.g. 3.8 points across 25 problems, if the teacher were simply multiplying to generate a 100 point scale).

3.2 Data Selection

In this study, we are seeking to understand the degree to which specific features of AAVE affect the grading patterns of LLMs. Therefore, we selected student-generated open-ended responses that contain features known to occur in AAVE. Because AAVE features can sometimes occur in other non-standard English varieties, we also ensured that these problems came from schools that were likely to serve a high population of AAVE speakers. That is, student-level data is not collected by ASSISTments, but we ensured that the students who produced these sentences were in schools with a high population of African American students to increase the odds that this dialect would be likely to be used in classroom

contexts. School-level data was obtained from the National Center for Education Statistics [22].

3.3 Producing Control Data

3.3.1 Preliminary Data Cleaning

The initial dataset contained 4,915 observations, where each row represented a problem. In ASSISTments, many problems include sub-problems that prompt students to answer conceptual questions, such as "why or why not?" or "explain your thinking." These conceptual sub-problems were represented in the initial dataset independently, without reference to the main problem. To ensure the LLM had sufficient context to accurately score these conceptual sub-problems, we manipulated the data to include the main problem along with all associated sub-problems. This step ensured that each observation represented the complete set of problem parts. Additional conceptual problems were removed due to not finding their parent problems. Lastly, we removed problems and student answers that contained viewing or uploading an image. After cleaning the data, there were 2,971 observations left.

Preliminary data identification was made by 3 members of the research team (the 1st, 3rd, and 4th authors), and involved selecting approximately 150 sentences for each feature. They did so by extracting problems that had word sequences likely to include a given construction (e.g., "you is" for subject/verb agreement leveling). After preliminary cleaning the data, we were left with 581 observations. The following count for each of the four features: 140 possible instances of negative concord, 148 instances of general s/v agreement leveling, 150 instances of preterit done, and 143 instances of habitual be.

3.3.2 Secondary Data Cleaning

Next, a secondary data cleaning process was conducted by hand. During this process, only sentences thought to contain these four grammatical features were checked by the 2nd author—who has extensive formal education and training specific to this task. The 2nd author removed sentences that did not contain these specific features from further analysis. This resulted in approximately 97 verified instances of negative concord, 106 instances of general s/v agreement leveling, 139 instances of preterit done, and 52 instances of habitual be.

Further inspection shows that these 394 responses were found in 349 ASSISTments problems from 75 distinct schools, with 86% of these schools located in the Southern United States.

Table 2. Number of Student Answers Containing a Given AAVE Grammatical Feature Before/After Data Cleaning

	Presented (n)	**Verified (n)**	**Verified (%)**
Negative Concord	140	97	69%
S/V Agreement Leveling	148	106	72%
Preterit Done	150	139	93%
Habitual Be	143	52	36%

3.3.3 Producing Control Data

Data cleaning was also required to produce the control data. In this process, a copy of each of the 394 selected sentences was made for the control group. Next, the 2nd author replaced the dialect feature targeted by this research question (and only the dialect feature targeted in this research) with a standard form. As Table 3 shows, in some cases, this resulted in a student response that used fully standard English (e.g., Examples B, I, and J).

However, in most cases there were still non-standard grammatical features in the control version of the student's response. For instance, Example A shows the shortening of the word "about" to "bout" which is common in many spoken English varieties. Likewise, Example C shows two patterns related that can be found in AAVE, "Why two of my answers are the same" (a pattern related to wh-question inversion strategies that differ across dialects; [16]) and "hudred," which could be related to a pronunciation pattern where nasal consonants are deleted.

Likewise, we did not correct for spelling or typos, and the sentences in this corpus show a range of patterns with respect to those issues as well. In other words, most of the control data still showed patterns that could have been marked for spelling or mechanics, even if the targeted feature was the only example of AAVE in the student's response.

Table 3. Examples of original student responses and their corresponding controls. Note the Ø symbol is used when an AAVE variable is deleted to create the standard (in contrast with times where it was replaced by another feature)

	Original	**Control**	**Remaining AAVE Features**	**Typos/Punctuation**
A	Its N because you is talking bout answers.	Its N because you are talking bout answers.	*bout* for *about*	*Its* for *It's*
B	No. Because you is adding 200 and then subtracting 100.	No. Because you are adding 200 and then subtracting 100.		Incomplete sentence
C	Why two of my answer is the same is because they is in the same family. hudred thousand.	Why two of my answer are the same is because they are in the same family. hudred thousand.	*Why two of my answers* for *The reason two of my answers*; *hudred* for *hundred*	Incomplete sentence
D	because there will be only one answer because i am in 6th and i ain't in no other	because there will be only one answer because i am in 6th and i am not in another grade that's why it is not a statistical question.		Incomplete sentence Capitalization/punctuation

	Original	Control	Remaining AAVE Features	Typos/Punctuation
	grade that's why it is not a statistical question.			
E	The zero's didn't have no effect on the mean, so they were irrelevant.	The zero's didn't have any effect on the mean, so they were irrelevant.		Zero's for zeros.
F	The hours that someone is there, no one can't pay no more than 12$ dollars.	The hours that someone is there, any one can pay Ø more than 12$ dollars.		Incomplete sentence
G	Kyle had 8 accounts in the bank when he went. He had borrowed a quarter from his uncle Ben to buy a gumball from the machine since he didn't have no more money in his accounts at the bank. What rational number represents the number of money he need to pay back his uncle Ben?	Kyle had 8 accounts in the bank when he went. He had borrowed a quarter from his uncle Ben to buy a gumball from the machine since he didn't have any more money in his accounts at the bank. What rational number represents the number of money he need to pay back his uncle Ben?	*Need* for *needs* (*note that this example of subject/verb agreement was not changed because this response was targeted for double negation)	
H	No but I didn't see no but I didn't see no line measurements	No but I didn't see any but I didn't see any line measurements		Punctuation, unnecessary repetition
I	We didn't get no cards!	We didn't get any cards!		
J	He done more than Andre.	He did more than Andre.		
K	when i done mine i got a quotient of 10	when i did mine i got a quotient of 10		Capitalization/punctuation

3.4 Prompt Engineering & GPT Settings

This study tests two prompts and three different temperature parameters. The design and selection of these drew heavily on previous research about prompting GPT to accommodate non-standard English [36, 37], but it also involved preliminary testing of prompts similar to those used in the Illustrative Math rubrics as well as testing of the temperature settings within GPT's API.

3.4.1 Preliminary Rubric Testing

Although state-of-the-art LLMs are becoming more powerful, writing effective prompts can be challenging [56]. In this study, we tested prompts that provided both (1) the context and task, (2) an established scoring rubric when asking the LLM to score each student response.

Specifically, we adapted a scoring rubric from Illustrative Mathematics (IM), a curriculum that leverages problem-solving tasks to help K–12 students learn math. The IM curriculum is found in a substantial proportion of the ASSISTments data, including in the questions that our data was drawn from. IM uses open-ended questions that fall under two types (*Conceptual Questions* and *Non-Conceptual Questions*), and student responses can also fall under two types (*Restricted Constructed Response* and *Extended Response*).

In this study, we tested the ability of the LLMs to implement IM's 5-point rubric, where Tier 4 represents the strongest answers and Tier 0 represents the weakest answers [24]. While the Tiers differ slightly for Conceptual and Non-Conceptual Questions, for simplicity and efficiency, we combined them to craft one common scoring rubric that can be used for both.

3.4.2 Preliminary Temperature Testing

Because previous research has found that both high and low temperature settings of GPT are better for different tasks [36, 37], we tested how high the temperature parameter could be set to before its creativity became untenable. Our results showed that temperatures above 0.5 were untenable, and so this study reports only on three settings at or below that mark: 0.0, 0.3, and 0.5.

3.4.3 Final Prompt Design

Once the instructions for scoring the mathematics part of the problem was defined, we also crafted text that could be used to provide instructions about non-standard dialect usage. In keeping with previous research on the use of LLMs in scoring non-standard English [36, 37], two types of prompts were designed for this study. Both used IM's standard rubric—already established—for grading open-ended conceptual problems in ASSISTments. The key difference between these two prompts was in the construction of definitions explicit to AAVE. As can be seen in Table 4, these included: (1)

Prompt 1, in which no dialect mentioned, and (2) **Prompt 2**, in which AAVE instructions are explicitly provided.

Table 4. Prompts used in this study, adapted from Illustrative Math's Rubric. Differences between the two prompts are underlined

Prompt 1 (no dialect mentioned)	**Prompt 2 (with AAVE instructions explicitly included)**
"I will provide middle school mathematics questions, a scoring rubric, and a middle school student's response. Based on the quality of the response and the provided rubric, please assign a score on a scale of 0 to 4, where 0 is the lowest score, and 4 is the highest. If the provided question has multiple parts, only consider the last part to score the student's response. The other parts are for your reference. Scoring Rubric: Use the following criteria to evaluate the responses: Tier 1 response: Work is complete and correct, with complete explanation or justification. Grade this response a 4. Tier 2 response: Work shows good conceptual understanding and mastery, with either minor errors or correct work with insufficient explanation or justification. Grade this response a 3. Tier 3 response: Work shows a developing but incomplete conceptual understanding, with significant errors. Grade this response a 2. Tier 4 response: Work includes major errors or omissions that demonstrate a lack of conceptual understanding and mastery. If the response shows effort, with major errors or omissions, grade the response with a 1. If the student did not try to answer the question, grade their response with a 0. **Question: " + cleaned.problem_bodies + " Student's Response: " + cleaned.answer_text + ". " "Score (0 to 4): [Provide the score here. Only provide the score]"**	"I will provide a 7th grade mathematics question, a scoring rubric and a 7th grade student's response. <u>These responses may contain grammatical features that are prevalent in African-American Vernacular English (AAVE). The features include: 1) habitual be, 2) preterit done, 3) double negatives, and 4) subject/verb agreement leveling. Please do not grade the students' problems based on their use of AAVE. You should treat both standard English and dialects the same in terms of scoring</u>. Based on the quality of the response and the provided rubric, please assign a score on a scale of from 0 to 4, where 0 is the lowest score and 4 is the highest. If the provided question has multiple parts, only consider the last part to score the student's response, the other parts are for your reference. Scoring Rubric: Use the following criteria to evaluate the responses: Tier 1 response: Work is complete and correct. Grade this response a 4. Tier 2 response: Work shows good conceptual understanding and mastery, with either minor errors or correct work with insufficient explanation or justification. Grade this response a 3. Tier 3 response: Work shows a developing but incomplete conceptual understanding, with significant errors. Grade this response a 2. Tier 4 response: Work includes major errors or omissions that demonstrate a lack of conceptual understanding and mastery. If the response shows effort, with major errors or omissions, grade the response with a 1. If the student did not try to answer the question, grade their response with a 0. **Question: " + cleaned.problem_bodies + " Student's Response: " + cleaned.answer_text + ". " "Score (0 to 4): [Provide the score here. Only provide the score]"**

3.5 Statistical Procedures

3.5.1 Descriptive

To compare how GPT graded students' original responses compared to the generated control group, we used R to generate data frames which simply contained the GPT grades for each of the conditions: Original response and Controlled response, Prompt 1 and Prompt 2, and the three temperatures, 0.0, 0.3, 0.5. We then calculated the mean and standard deviation for each condition, as well as reported the number of data points to have an understanding of the distribution of data across conditions and to determine if any additional cleaning was required.

3.5.2 ANOVA

To determine the degree to which the various temperature settings might be affecting the scores in our data, we applied an ANOVA analysis to scores to the four prompt and response conditions, comparing the temperatures of each. Because each condition had data for 3 different temperatures, we ran an ANOVA which told us if there was significant variance across the three. Once we had the ANOVA results, we could continue with the data analysis by selecting a single temperature.

3.5.3 Paired T-Test

To control for the multiple conditions we ran four paired Student T-tests as follows: (1) Original vs. Control Data with Prompt 1, (2) Original vs. Control Data with Prompt 2, (3) Prompt 1 vs. Prompt 2 with the original data, and (4) Prompt 1 vs. Prompt 2 with the control data. The results of these T-tests tell us if there is any statistical difference between the two compared conditions. We did not compare any other combinations of prompt and response data (e.g., Original with Prompt 1 vs. Control with Prompt 2) as the four conditions listed above are sufficient for answering our major research questions (i.e., RQ2 and RQ3).

4. RESULTS

4.1 Descriptive Statistics

To begin our analysis of how GPT grades students' original responses compared to the generated control group, we calculated the mean and standard deviation for each condition and temperature (Original vs Control, Prompt 1 vs Prompt 2, and Temperatures 0.0, 0.3, and 0.5). As Table 5 shows, all standard deviations were low (below 1), indicating little variance in the data, and confirming the absence of outliers. Additionally, the means across conditions, but within temperatures, were consistent with little variation (see Table 6, next section for ANOVA results).

Table 5. Descriptive Statistics

GPT Settings		Original			Control		
Prompt	**Temp**	**Mean**	**SD**	**n**	**Mean**	**SD**	**n**
Prompt 1	**0.0**	1.35	0.81	394	1.38	0.83	394
	0.3	1.37	0.81	394	1.38	0.86	394
	0.5	1.36	0.82	394	1.39	0.87	394
Prompt 2	**0.0**	1.68	0.95	394	1.63	0.93	394
	0.3	1.70	0.96	394	1.62	0.93	394
	0.5	1.70	0.95	394	1.67	0.93	394

4.2 ANOVA

To determine the degree to which the various temperature settings might be affecting the scores in our data, we applied an ANOVA analysis to scores from all conditions (original vs control data, Prompt 1 vs Prompt 2, temperatures of 0.0, 0.3, and 0.5). As Table 6 shows, there were no significant differences in temperature (all p values are greater than 0.05), and so subsequent analyses were run exclusively for temperature 0.5.

Table 6. ANOVA Testing Temperature Settings

Condition	**df (within groups)**	**df (between groups)**	**F**	**Sum Sq**	**p**
Original, Prompt 1	1272	2	0.058	0.1	0.944
Original, Prompt 2	1272	2	0.079	0.1	0.924
Control, Prompt 1	1272	2	0.003	0.0	0.997
Control, Prompt 2	1272	2	0.369	0.6	0.692

4.3 T-Tests

T-Tests were run for four comparisons. First, we compared the original data to the control data using the scores from Prompt 1 (no AAVE). Then we compared the original data to the control data using the scores from Prompt 2. We then compared Prompt 1's scoring of the original data to Prompt 2's scoring of the original data. Finally, we compared Prompt 1's versus Prompt 2's scoring of the student responses. An overview of these results is shown in Table 7.

Table 7. T-test Results; significant results are in grayscale

	Mean Diff.	**CI**	**p**
Original vs. Control Data (Prompt 1)	-0.02	-0.13 – 0.09	0.713
Original vs. Control Data (Prompt 2)	0.03	-0.09 – 0.16	0.629
Prompt 1 vs Prompt 2 (Original Data)	0.34	0.29 – 0.39	2.2e-16
Prompt 1 vs Prompt 2 (Control Data)	0.26	0.23 – 0.28	2.2e-16

4.3.1 Scores for Original vs. Control (Prompt 1)

Recall that Prompt 1 had no explicit instructions regarding dialect, and that our two data sets varied in the amount of AAVE present in each problem. Our first T-test compares the original data (higher AAVE) to the control data (with one of the four grammatical features standardized). This establishes a baseline for our other results and allows us to test the degree to which minor changes towards a more standard dialect pattern would improve the scores without prompt engineering efforts.

The results of the paired T-test indicate that we can make no claim about the difference in score between the original response and control response (Table 7). In other words, simply changing a single grammatical feature in each problem answer was not sufficient for statistically raising the scores ($p \leq 0.713$).

4.3.2 Scores for Original vs. Control (Prompt 2)

Our second T-test compares the scores for the two data sets when both were evaluated using Prompt 2--which provided explicit instructions not to penalize grammatical features of AAVE. The results of the T-test imply that no claims towards the difference of control vs original responses can be made ($p \leq 0.629$). This is consistent with our findings from section *4.3.1*.

4.3.3 Scores from Prompt 1 vs. Prompt 2 (Original)

Our next T-test compares the scores of the original data when graded using the prompt that does not give any instruction about dialect (Prompt 1) to the scores it provides when it is given explicit instructions not to penalize grammatical features (Prompt 2). This allows us to test the degree to which prompt engineering efforts might be used to ameliorate any biases toward non-standard dialect patterns.

The results show that there is a statistically significant difference between the scores of the two prompts with prompt 2 producing higher scores. In other words, we were able to make a significant difference in the scores from prompt engineering alone (Mean difference=0.34; $p \leq 0.000$). Additionally, the confidence interval (CI = 0.29 - 0.39) indicates that with 95% confidence the difference between the two prompts will fall within that range.

4.3.4 Scores for Prompt 1 vs. Prompt 2 (Control)

When comparing the scores of the controlled data with both Prompt 1 and Prompt 2, we found that there is a statistically significant difference between GPT scores of each prompt (Mean difference=0.26; ($p \leq 0.000$) The positive mean difference indicates that Prompt 2 produces higher scores from GPT. These results are in line with those we presented for Prompt 1 (Section 4.3.3) but shows that the effect is still present even when there are fewer AAVE features present in the data. $p \leq 0.000$).

5. DISCUSSION

5.1 Overview of Results

This study investigated GPT-4o's ability to score open-ended mathematical responses containing AAVE features, such as negative concord, general subject/verb agreement leveling, habitual be, and preterit done. We first created a control dataset in which each AAVE feature in the response was replaced with its SAE equivalence (see Table 3). Each response was selected because it had one of four specific dialect features (double negation, subject/verb agreement leveling, preterit done, or habitual be), but it may have had other instances of non-standard grammar as well.

To determine which temperature setting to use (**RQ1**), we explored three different temperature settings (0, 0.3, and 0.5) to determine which setting influenced grading scores. As shown in Table 6, all three settings performed similarly the same and showed no significant variation in scores. Because there were no differences, we used temperature setting 0.5 to address RQ2 and RQ3.

For **RQ2**, we test the degree to which standardizing students' responses might change the way GPT scores those answers. To do so, we compared the original data to the data where one of the four dialect features had been removed, but we continued to compare the performance of Prompt 1 and Prompt 2. Results from sections 4.3.1 and 4.3.2 indicate that explicitly specifying or omitting AAVE features did not lead to statistically significant differences in scoring. This suggests that GPT-4o scored response in the original and control datasets relatively similar. In other words, altering a single grammatical feature in each response did not significantly impact GPT's scoring.

Although these results show no evidence of bias against responses that contain AAVE features, these were messy data and further testing is needed to determine whether or not these results would hold if the control group had gone through further standardization changes.

To address **RQ3**, we tested the degree to which prompt-engineering could mitigate potential scoring biases. To do so, we applied two prompts to both the original dataset (with AAVE features) and the control dataset (with SAE equivalents). Results from sections 4.3.3 and 4.3.4 indicate that explicitly indicating the presence of AAVE features (Prompt 2) led to statistically higher scores compared to omitting the presence of AAVE features (Prompt 1). Prompt 2 produced higher scores for both comparisons, suggesting that we should continue to test prompt engineering strategies that might be used to mitigate these biases. Readers should note that this will likely include some degree of testing with synthetic data, as [36, 37] have shown that the sorts of human labels typically used to train our models likely also contain significant biases towards non-standard dialects.

5.2 Limitations & Future Work

This study presents what we believe to be the first examination of the effects of non-standard dialect patterns on the automated scoring of math problems by an LLM, but this study is not without limitations. For example, while we have chosen to focus on naturalistic data that was retrieved from the ASSISTments system, this does mean that our data is not well controlled for several factors. These responses represent a range of differences in content and length. They also contain differences related to other nonstandard grammar patterns (as the control data only removed one instance per response) as well as in spelling and punctuation errors.

For these reasons, and because of the low number of instances in some categories (e.g., *habitual be, N=52*), we cannot say specifically which grammatical features might most influence GPT's scoring patterns. Since it is unclear how GPT (and other LLMs) has sought to decrease biases in these models, it is difficult to suggest exact steps forward for their ability to respond more effectively to students from non-standard dialect backgrounds. It seems likely that features like *habitual be* would have fewer instances represented in the training data of most LLMs than grammatical features like *double negation*, which are both more common in other kinds of non-standard dialects and typical in other languages.

Future work should expand the scale of these efforts within naturalistic data, so that we are able to capture the kinds of messy data that young students often produce. However, testing these issues with fully synthetic data could help us to pinpoint the kinds of grammatical variation that is likely to cause these LLMS to behave differently, which could also provide better information for prompt engineering and other efforts to ensure that LLMS can effectively handle a range of different dialect patterns.

5.3 Potential Implications

Despite the limitations in our work, we show results that might have been more consequential had the data been more highly controlled. For example, if we had removed all grammatical features of AAVE from the control data (as opposed to one of the four features targeted for analysis in this study), we might have seen statistically significant differences in the scoring of those samples within at least Prompt 1.

That said, the scoring differences between Prompt 1 and Prompt 2 are consequential in and of themselves, particularly given the concerns that [36, 37] have raised about GPT's ability to pinpoint the kind of grammatical difference that it is reacting to. That is, these studies showed that GPT reacted more strongly to AAVE than it did to a bad essay (with poor syntax and weak lexical choices), but they also showed that it could not correctly identify the parts of the speech that it considered problematic in those sentences.

5.3.1 Effects of Different Grade Calculations

In this study, we show that by changing the prompt to explicitly mention AAVE features and to instruct GPT to ignore them, we can raise the grades in both the original and the control data. Table 8 shows the mean differences we reported upon above as well as the effect that these might have if the 4-point scale were translated to a percentage-based grading system.

Table 8. Approximate reductions in student grades

Prompt 1 vs Prompt 2	Original	Control	Diff
Effect on 4pt-scale	-0.34	-0.26	-.08
Effect on Percentage Scale	-8.50%	-5.5%	-3%

For situations in which a teacher calculates the grade on a 4-point grading scale, a student who is otherwise answering perfectly is likely to receive a small but noticeable reduction in their grade. How much that difference would be is somewhat difficult to say given the parameters of this study's design. These results show that even when we have removed one instance of AAVE grammatical features from the data (i.e., the Control data), Prompt 2 raises the grade by a quarter of a letter grade. That effect is stronger for the Original data, where one of the four grammatical features we identified was known to be present.

In other words, even though the difference between the grades for these two data sets (Original vs. Control) are not significantly different, at least part of the change in the scoring of the Original data is likely related to the fact that either (a) other features of AAVE were present, or (b) the system was treating typos as the same as dialect, which it has been shown to do in the past [36, 37]. Future work should also test several versions of the data (AAVE only, typos only, combined, etc.) so that we can better understand which features of student answers are most likely to impact these scores.

5.3.2 Impact of Grade Calculations on Students

The impact of this prompt design might be relatively small for a student who is otherwise performing well, but even a one quarter reduction on a 4-point scale (the effect on the Control data) could have more serious ramifications for a student who is already struggling, including the possibility of changing letter grades or even failing a class. This effect could be magnified if the teacher were using a percentage-based grading system. Here, a 0.34 reduction on a 4-point problem would result in an 8.5% reduction of a percentage grade. For a child who was otherwise performing perfectly (e.g., a drop from 100% to 91.5%), might not receive any major penalties, but it could remove a high school student out of contention for top class rankings, which can improve admission and scholarship odds for college. Such a reduction could also prevent an otherwise strong performer from being considered for an honors math class (e.g., 90% to 82%) or worse, retained for the year (e.g., 73% to 65%).

Although the improvements in the scores for the Original data suggest that Prompt 2 could be mitigating more than just the effect of dialect, we could also consider what the effects might be on a student if we assumed (falsely) that the Control data contained no AAVE features. If this were true, the difference between the Prompt 2 effect on the Original data (-0.34) versus its effect on the Control data (-0.26) still show a one-fifth letter grade penalty on a 4-point scale and a 3% (one-third letter grade) penalty on a percentage scale.

Moreover, even if the differences are small, they demonstrate that—as with the essay scoring research--the automated scoring is picking up on something other than the quality of the work in that is targeted for evaluation. Given such potential consequences, future research should consider how these patterns might manifest in other contexts and learning domains, but also what additional steps could be used to mitigate this differential treatment in large language models.

6. CONCLUSIONS

The need for research on how LLMs respond to dialect differences is important beyond the scope of education research [4, 20, 27, 38], but its utility takes on particular importance when scoring and responding to student work. If LLMs like GPT are significantly affected by the presence of non-standard grammatical features, this will complicate the ability to provide the kind of automated feedback necessary to improve student learning.

We also note that, within education, these issues extend beyond dialect to those related to students with speech and language issues that can also produce patterns that are not a regular part of LLM training data [1, 23]. Although most of this research does not appear to be taking place in educational contexts, there is a history of looking for these kinds of differences available in the literature that could help pave the way forward [49, 57]. More research is needed to determine the best approaches for mitigating the difficiencies caused by the lack of training data in these models.

This study differs from previous work on automated essay scoring (AES) by focusing specifically on automated content scoring (ACS). In ACS, the score is based on the content of the response, regardless of grammar or spelling errors. In addition, this study explored the degree to which biases toward non-standard dialects might interfere with an LLM's ability to provide appropriate scoring in a computer-based learning platform for math. In doing so, we hope that we have demonstrated the need for greater attention to these issues in STEM domains. Not only could they help to ensure that we are providing fairer and more accurate evaluations to students, but they could tell us more about LLMs and how their internal structures perform more generally.

7. REFERENCES

[1] Addy, T., Kang, T., Laquintano, T., & Dietrich, V. (2023). Who Benefits and Who is Excluded?: Transformative Learning, Equity, and Generative Artificial Intelligence. *Journal of Transformative Learning*, *10*(2), 92-103.

[2] ASSISTMENTS. (2025). *ASSISTments*. Retrieved February 15, 2025, from https://new.assistments.org/

[3] Baron, D. E. (1975). Non-standard English, composition, and the academic establishment. College English, 37(2), 176-183.

[4] Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021, March). On the dangers of stochastic parrots: Can language models be too big? 🦜. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency* (pp. 610-623).

[5] Benjamin, R. (2023). Race after technology: Abolitionist tools for the new Jim code. Polity Press.

[6] Charity Hudley, A. H.., Mallinson, C., & Bucholtz, M. (2020). Toward racial justice in linguistics: Interdisciplinary insights into theorizing race in the discipline and diversifying the profession. Language, 96(4), e200-e235.

[7] Doewes, A., Kurdhi, N. A., & Saxena, A. (2023). Evaluating Quadratic Weighted Kappa as the Standard Performance Metric for Automated Essay Scoring. In Proceedings of the 16th International Conference on Educational Data Mining (pp. 145-156).

[8] Eisikovits, E. (1991). Variation in subject-verb agreement in Inner Sydney English. *English around the world: Sociolinguistic perspectives*, 235-255.

[9] Fasold, R. W. (1971). What Can an English Teacher Do About Nonstandard Dialect?. English Record, 21(4), 82-91.

[10] Fasold, R. W. (1971). What Can an English Teacher Do About Nonstandard Dialect?. *English Record*, *21*(4), 82-91.

[11] Fought, C. (2006). Language and Ethnicity. Cambridge University Press.

[12] Gabbay, H., & Cohen, A. (2024, July). Combining LLM-generated and test-based feedback in a MOOC for programming. In Proceedings of the Eleventh ACM Conference on Learning@ Scale (pp. 177-187).

[13] Goldshtein, M., Ocumpaugh, J., Potter, A., & Roscoe, R. D. (2024, June). The Social Consequences of Language Technologies and Their Underlying Language Ideologies. In International Conference on Human-Computer Interaction (pp. 271-290). Cham: Springer Nature Switzerland.

[14] Green, L. (2017). Beyond lists of differences to accurate descriptions. In *Data Collection in Sociolinguistics* (pp. 281-284). Routledge.

[15] Green, L. J. (2002). *African American English: a linguistic introduction*. Cambridge University Press.

[16] Green, L., Wyatt, T. A., & Lopez, Q. (2007). Event Arguments and'Be'in Child African American English.

[17] Hazen, K. (2000). Subject-verb concord in a postinsular dialect: The gradual persistence of dialect patterning. Journal of English Linguistics, 28(2), 127-144.

[18] Heffernan, N. T., & Heffernan, C. L. (2014). The ASSISTments ecosystem: Building a platform that brings scientists and teachers together for minimally invasive research on human learning and teaching. *International Journal of Artificial Intelligence in Education*, *24*, 470-497.

[19] Hickey, R. (2007). Tracking dialect history: a corpus of Irish English. In *Creating and digitizing language corpora: Volume 2: Diachronic Databases* (pp. 105-126). London: Palgrave Macmillan UK.

[20] Hofmann, V., Kalluri, P. R., Jurafsky, D., & King, S. (2024). Dialect prejudice predicts AI decisions about people's character, employability, and criminality. arXiv preprint arXiv:2403.00742

[21] Hollie, S. (2017). Culturally and linguistically responsive teaching and learning: Classroom practices for student success. Teacher Created Materials.

[22] Hussar, W. J., & Bailey, T. M. (2011). Projections of Education Statistics to 2020. NCES 2011-026. *National Center for Education Statistics*.

[23] Hutchinson, B., Prabhakaran, V., Denton, E., Webster, K., Zhong, Y., & Denuyl, S. (2020). Social biases in NLP models as barriers for persons with disabilities. *arXiv preprint arXiv:2005.00813*.

[24] Illustrative Mathematics Summative Assessments—Teachers | IM Demo. (n.d.). Retrieved February 20, 2025, from https://curriculum.illustrativemathematics.org/HS/teachers/summative_assessments.html

[25] Jackson, J. E., & Green, L. (2005). Tense and aspectual be in child African American English. Perspectives on aspect, 233-250.

[26] Jankowski, B. L., & Tagliamonte, S. A. (2022). He come out and give me a beer but he never seen the bear: Vernacular preterites in Ontario dialects. English World-Wide, 43(3), 267-296.

[27] Kirk, H. R., Vidgen, B., Röttger, P., & Hale, S. A. (2024). The benefits, risks and bounds of personalizing the alignment of large language models to individuals. *Nature Machine Intelligence*, *6*(4), 383-392.

[28] Labov, W. (1969). A Study of Non-Standard English.

[29] Labov, W. (1972). Academic ignorance and black intelligence.

[30] Labov, W. (2000). The logic of non-standard English (1969). *L. Burke, T. Crowley and A. Girvin, The Routledge Language and Cultural Theory Reader*, 456-466.

[31] Meyer, J., Jansen, T., Schiller, R., Liebenow, L. W., Steinbach, M., Horbach, A., & Fleckenstein, J. (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, 100199.

[32] Mizumoto, A., & Eguchi, M. (2023). Exploring the potential of using an AI language model for automated essay scoring. Research Methods in Applied Linguistics, 2(2), 100050.

[33] Mosher, M., Dieker, L., & Hines, R. (2024). The Past, present, and future use of artificial intelligence in teacher education. Journal of Special Education Preparation, 4(2), 6-17.

[34] Mufwene, S. S. (2008). What is African American English?. In *Sociocultural and historical contexts of African American English* (pp. 21-52). John Benjamins Publishing Company.

[35] Nesbitt, J. (2022). Writing while Black: African American vernacular English (AAVE) and perceived writing performance (Doctoral dissertation, Dissertation, James Madison University, Harrisonburg, VA]. https://commons. lib. jmu. edu/cgi/viewcontent. cgi).

[36] Ocumpaugh, J., Liu, X., & Zambrano, A. F. (2025, March). Language Models and Dialect Differences. In *Proceedings of the 15th International Learning Analytics and Knowledge Conference* (pp. 204-215).

[37] Ocumpaugh et al., (2025b) Assessment of Training Data for Asset-based Technology. Paper accepted to the 2025 Winter Meeting of the Society for Text & Discourse.

[38] Ostrand, R., & Berger, S. E. (2024, July). Humans linguistically align to their conversational partners, and language models should too. In *ICML 2024 Workshop on LLMs and Cognition*.

[39] Porwal, R., Rozet, A., Houck, P., Gowda, J., Moeller, S., & Tang, K. (2025). Analysis of LLM as a grammatical feature tagger for African American English. arXiv preprint arXiv:2502.06004.

[40] Pride, J. (1974). Deficit—Difference controversy. Archivum Linguisticum, 5, 35.

[41] Priest, K. (2018). Decreasing Teacher Burnout Through Teaching Effective and Efficient Grading Strategies.

[42] Pullum, G. K. (1999). African American Vernacular English is not standard English with mistakes. The workings of language: From prescriptions to perspectives, 59-66.

[43] Quinn, H. (2011). Variation in New Zealand English syntax and morphology. In New Zealand English (pp. 173-197). John Benjamins Publishing Company.

[44] Ramesh, D., & Sanampudi, S. K. (2022). An automated essay scoring systems: a systematic literature review. Artificial Intelligence Review, 55(3), 2495-2527.

[45] Schmucker, R., Xia, M., Azaria, A., & Mitchell, T. (2024). Ruffle & Riley: Insights from designing and evaluating a large language model-based conversational tutoring system. In International Conference on Artificial Intelligence in Education (pp. 75-90). Cham: Springer Nature Switzerland.

[46] Shakya, A., Rus, V., & Venugopal, D. (2023). Scalable and Equitable Math Problem Solving Strategy Prediction in Big Educational Data. In Proceedings of the 16th International Conference on Educational Data Mining (pp. 169-180).

[47] Squires, L. (2014). Processing, evaluation, knowledge: Testing the perception of English subject–verb agreement variation. Journal of English Linguistics, 42(2), 144-172.

[48] Starace, G., Papakostas, K., Choenni, R., Panagiotopoulos, A., Rosati, M., Leidinger, A., & Shutova, E. (2023). Probing LLMs for Joint Encoding of Linguistic Categories. arXiv preprint arXiv:2310.18696.

[49] Voigt, R., Jurgens, D., Prabhakaran, V., Jurafsky, D., & Tsvetkov, Y. (2018, May). RtGender: A corpus for studying differential responses to gender. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

[50] Winford, D. (1998). On the origins of African American vernacular English—A creolist perspective: Part II: Linguistic features. Diachronica, 15(1), 99-154.

[51] Wolfram, W. (1969). Sociolinguistic Premises and the Nature of Nonstandard Dialects.

[52] Wolfram, W., & Kohn, M. E. (2015). Regionality in the development of African American English. The Oxford Handbook of African American Language, 140-160.

[53] Wolfram, W., & Thomas, E. (2008). The Development of African American English. John Wiley & Sons.

[54] Wolfram, W. (2004). The grammar of urban African American vernacular English. *Handbook of varieties of English*, *2*, 111-32.

[55] Xiao, C., Ma, W., Xu, S. X., Zhang, K., Wang, Y., & Fu, Q. (2024). From Automation to Augmentation: Large Language Models Elevating Essay Scoring Landscape. arXiv preprint arXiv:2401.06431.

[56] Zamfirescu-Pereira, J. D., Wong, R. Y., Hartmann, B., & Yang, Q. (2023, April). Why Johnny can't prompt: how non-AI experts try (and fail) to design LLM prompts. In Proceedings of the 2023 CHI conference on human factors in computing systems (pp. 1-21).

[57] Ziems, C., Held, W., Yang, J., Dhamala, J., Gupta, R., & Yang, D. (2022). Multi-VALUE: A framework for cross-dialectal English NLP. *arXiv preprint arXiv:2212.08011*.