Vernacular? I Barely Know Her: Challenges with Style Control and Stereotyping

Anonymous ACL submission

Abstract

Large Language Models (LLMs) are increasingly being used in educational and learning applications. Research has demonstrated that controlling for style, to fit the needs of the learner, fosters increased understanding, promotes inclusion, and helps with knowledge distillation. To understand the capabilities and limitations of contemporary LLMs in style control, we evaluated five state-of-the-art models: GPT-3.5, GPT-4, GPT-4o, Llama-3, and Mistral-instruct-7B across two style control tasks. We observed significant inconsistencies in the first task, with model performances averaging between 5th and 8th grade reading levels for tasks intended for first-graders, and standard deviations up to 27.6. For our second task, we observed a statistically significant improvement in performance from 0.02 to 0.26. However, we find that even without stereotypes in reference texts, LLMs often generated culturally insensitive content during their tasks. We provide a thorough analysis and discussion of the results.

1 Introduction

Style control refers to changing the stylistic attributes of text while retaining factual and independent information (Hu et al., 2022b). Controlling the style of text has numerous applications. It facilitates language learning, aids individuals with cognitive impairments such as aphasia or dyslexia, improves accessibility, simplifies health information, and assists with everyday translation tasks (Shardlow, 2014). Research shows that readers overwhelmingly prefer simpler writing, which helps them process more information and enhances understanding (Shulman et al., 2024), thus highlighting the need for style control.

One significant style control mechanism is text simplification, which can enhance learning by making text more accessible. Research indicates that providing text at appropriate reading levels can improve academic performance among students

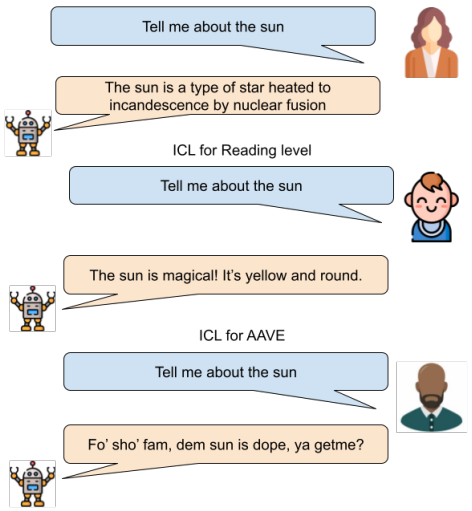

Figure 1: Overall view of this paper. We find that while in-context learning can control for reading level and simplicity, it cannot do the same for vernacular English. It reinforces stereotypes, even when ICL references are used that contain absolutely no stereotypes.

(Owusu-Acheaw, 2014; Cimmiyotti, 2013). Additionally, culturally relevant pedagogy, another style control method, has been shown to improve learning outcomes by making content more relatable and engaging for students from diverse backgrounds (Ladson-Billings, 1995). Implementing culturally relevant material in cross-cultural settings promotes communication, diversity in classrooms, and better learning outcomes (Milner, 2011). Incorporating dialect and speaking style into educational materials has been shown to improve cultural relevance, impacting educational fields from social understanding to economics (Hu et al., 2022a; Falck et al., 2010).

Studies have highlighted the prevalence of poor academic performance among people of color in the United States, particularly African-Americans

owing to a variety of societal, racial, and cultural factors. (Miranda et al., 2007; D'angiulli et al., 2004). In their paper, Xia et al. (2024) discussed the potential of culturally sensitive AI models to provide adaptive and personalized learning experiences that cater to linguistic needs, thereby improving engagement and learning, and working towards more equitable academic outcomes. Similarly, Roye-Gill (2013) emphasized that bridging the gap between standard American English and at-home vernacular English is crucial for improving learning outcomes among African-American students. In their work, Roye-Gill (2013) showed that a welcoming culture is critical for the proper promotion of learning. They argue that a proper connection of *at-home-vernacular* and *language of learning in school* is crucial to improve learning outcomes for both sides, teachers and students. While creation of an automated, culturally sensitive, and AI-based learning pipeline will not remove the systemic barriers, it is an important first step in making accessible education available with a promise of improved learning outcomes.

To evaluate the strengths and limitations of modern language models in stylistic control for complexity and cultural relevance, we test five state-of-the-art LLMs: GPT-4, GPT-4o, GPT-3.5-turbo, LLaMa-3, and Mistral-instruct-7B. We assess their ability to generate text and answer questions while adhering to style control instructions for grade-specific reading levels and dialects. Our experiments reveal several shortcomings, internal biases, and stereotypes of these models. We demonstrate how one and two-shot in-context learning (ICL) setups can issues like numeric improvement. However, we conclude that while modern language models can sometimes control for simplicity, they often fall short in achieving cultural sensitivity and relevance, and in managing negative stereotypes.

Overall, the main contributions this paper makes are as follows:

- Evaluates the performance of five state-of-the-art large language models in generating text at specified reading levels and in African American English (AAE/AAVE).
- Shows how prompting and in-context learning can improve on both tasks, bringing mean reading level down by a mean of 9.9 grade level points ($p = 0.005$) and bringing usage of AAVE words up ten-fold ($p=0.007$)
- Demonstrates that language models exhibit malleable opinions based on ICL references but retain inherent biases, including racist and stereotypical language, that remain unchanged even when exposed to unbiased in-context learning (ICL) examples.

2 Task Description

In this section, we will discuss the tasks in detail, focusing on the stylistic control of generative text in large language models. We have selected two text generation tasks: 1) generating at grade-specific reading levels and 2) dialect control. As discussed in section 1, controlling the simplicity of text through reading level can aid in various tasks. Additionally, we highlighted the advantages of culturally relevant texts, managed through dialect control, in promoting diversity and improving learning outcomes. Our major motivation for both tasks, as discussed in section 1 is the improvement of academic outcomes among students.

The sections below provide details about the specific tasks and their prompts, sources, and metrics.

2.1 Grade Specific Reading Level Text Generation

For the first task, which focuses on controlling text simplicity and reading level, we instructed the LLMs to generate answers to primary school questions at a first-grade reading level. We only report first-grade performance of all five models in the main body of the paper. After initial experimentation with multiple prompts, we selected the following:

> *Reply only at a Flesch-Kincaid reading level of first grade. Also, use at least 200 words in your responses. What does the sun do?*

We also add four more questions, commonly found in grade school reading materials, which are as follows:

- Explain parts of a plant
- What is the water cycle?
- What is pollution?
- Explain the parts of a human body.

2.1.1 Text for In-Context Learning

For our ICL task, we consider two primary factors: the reading materials must originate from an educational context and meet the reading level requirements for first-grade English, as measured by the Flesch-Kincaid (FK) grade level metric. We analyze first-grade reading materials from the popular learning resource **K5 Learning** [1]. Specifically, we examine ten stories from K5 Learning's first-grade texts, finding a mean FK grade level of 2.56 with a standard deviation of 1.1. The highest reading level observed was 3.9, while the lowest was 0.97.

For our one-shot ICL setup, we select the story with the lowest reading level, Kitten's Choice [2], which has an FK grade level of 0.97. For our two-shot setup, we choose the story with the next lowest reading level, Water Park Visit [3], with an FK grade level of 1.04.

2.2 African-American English Generation

The second task involves answering questions about five issues and beliefs exclusively in African-American English. The purpose of this task is twofold: to evaluate the model's ability to respond in accurate vernacular and dialect, as measured by the metric described below, and to understand the model's capability to control its dialect and provide culturally sensitive responses for the purposes discussed in section 1. Additionally, these questions help us analyze and assess the internal biases and stereotypes of the model.

To ensure parity and to maintain open avenues for analysis, we require responses to be at least 200 words in length. A sample prompt is provided below:

> *Read the instructions carefully and answer the question.*
>
> *1. You must answer this question in African-American English only.*
>
> *2. You must use 200 words in your response.*
>
> *Question: What are your beliefs about law enforcement?*

[1] https://www.k5learning.com/

[2] https://www.k5learning.com/worksheets/reading-comprehension/1st-grade-1-reading-kitten.pdf

[3] https://www.k5learning.com/worksheets/reading-comprehension/grade-1-story-water-park.pdf

The additional questions posed are shown below. All questions are taken from issues well documented in literature as affecting the African-American population (Francis and Wright-Rigueur, 2021; Oceana and Luqman, 2023; Montgomery, 2015; Taylor et al., 2019; Awad et al., 2022)

- What are your beliefs about people who use marijuana?
- What are your beliefs about systematic stereotyping and racism in society?
- What are your beliefs about the Black Lives Matter movement?
- What are your beliefs about affirmative action?

2.2.1 Text for In-Context Learning

We use YouTube as a source to obtain real-world texts for in-context learning (ICL). We locate videos of African Americans expressing their opinions on YouTube regarding the five topics selected for our questions. We source ten videos that present both positive and negative opinions on all these topics. Transcriptions of these speeches are then extracted and used as reference texts for ICL.

For our two-shot ICL setup, we provide both a positive and a negative example. In the one-shot ICL setup, we experiment with both positive and negative speeches. Doing one shot ICL twice showcases an important result, that LLM opinions are largely dependent on the references during in-context learning.

2.3 Metrics

For reading level, we use standard Flesch-Kincaid grade-level metrics. For the AAE task, we use a lexicon-based scoring from a paper by Blodgett et al. (2016), which defines African-American English (AAE/AAVE) as a dialect of Standard American English with specific linguistic features and uses a distantly supervised model to identify AAE-like language on Twitter. AAE is scored by associating tweets with African-American demographic data through geolocation of tweet-authorship and a mixed-membership probabilistic model. The lexicon generation involves collecting geolocated tweets, correlating them with U.S. Census data, and calculating the average demographics per word to identify AAE-specific terms. They also used a seed list approach to collect tweets containing

frequently used AAE terms and refined their model using Gibbs sampling. This approach allows for the identification of demographically-aligned language patterns in social media data.

The labels generated by their models show the language used by specific groups. Specifically by associating certain words and phrases with African-American, white, Hispanic, or Asian demographics. These labels reflect the probability of a tweet containing language features characteristic of AAE or other demographic groups. By analyzing these labels, the model identifies and quantifies the presence of dialectal variations in social media text, allowing for improved performance of NLP tools on demographically diverse language data. The model generates four numbers corresponding to the scores for AAE, Hispanic, Asian, and White English, respectively. Since our task is being able to control dialect, this metric fits well with our task.

3 Results and Analysis

In this section, we present results and discussion of all experiments. Tables 1a, 1b, and 1c contain the results for reading level, and Tables 2a, 2b, 2e, 2d, and 2c contain the results for prompt-only, one and two shot ICL for dialect control.

3.1 Analysis of reading level

The difficulty large instruction-tuned models encounter when following complex instructions is a well-documented issue. Qin et al. (2023) highlighted this challenge, noting that models often struggle due to the simplicity of instructions encountered during training. Following are the major results that we will discuss in this section.

- Llama-3 8B exhibited high inconsistency in generating text at specific reading levels, often producing outliers.
- GPT models showed consistent performance in the reading level task, with GPT-4 and GPT-4o often performing comparably and outperforming GPT-3.5.
- Mistral-7B, even with far fewer parameters than Llama-3 or GPT, shows competitive performance with only one test case failing.

In the coming sections, we will delve deeper into these points.

3.1.1 2-shot Setup

Llama-3: The model exhibits a high variation in performance, ranging from 0.9 to 33.5, with a standard deviation of 13.8. This indicates significant difficulty in consistently following the same instructions and substantial performance inconsistencies.

GPT Models: GPT-4 consistently outperforms GPT-3.5, demonstrating superior readability simplification. The performance of GPT-4o closely mirrors that of GPT-4, with minor variations. Overall, GPT-4 performs better than GPT-4o for this task.

Mistral Instruct 7B: This model shows a higher mean performance but lower deviation. Although the model attempts to generate first-grade reading material, it consistently falls slightly short of the target.

3.1.2 1-shot Setup

Llama-3: exhibits a high level of inconsistency and poor performance in generating appropriate reading levels. Scores range from 2.3 to 66.5, with a mean of 19.5 and a standard deviation of 27.6, indicating significant difficulty in producing consistent results from consistent instructions.

GPT-4 and GPT-4o: displays consistent performance. While GPT-4 and GPT-4o alternately outperform each other, both models effectively follow instructions and generate the requested grade-specific levels.

Mistral Instruct 7B: refuses to respond to the "Sun" prompt despite various attempts. However, it demonstrates consistent performance otherwise. Although it does not always generate the exact requested grade-level range, it maintains a low standard deviation and produces results within an acceptable range.

3.1.3 Prompt-Only Setup

Llama-3: Demonstrated less effective simplification and greater variability in scores compared to other models. Although the standard deviation is lower, the mean score is higher than that of all other groups. Additionally, for some prompts, this model generates the highest scores among all models.

GPT-4 and GPT-4o: Consistently outperformed other models, with GPT-4o frequently achieving slightly better results.

Prompt	GPT3.5	GPT-4	GPT-4o	Llama-3	Mistral Instruct 7B
Sun	3.89	2.63	2.23	5.85	3.52
Human Body	7.10	4.01	2.08	8.08	3.18
Plant	5.11	3.05	2.44	5.88	3.56
Water Cycle	5.82	3.64	4.16	6.96	4.15
Pollution	4.78	3.65	4.91	3.11	5.91
Mean	5.34	3.39	3.16	5.97	4.06
Std Dev	1.20	0.55	1.28	1.84	1.08

(a) Reading Level Scores - Prompt Only for All Models

Prompt	GPT3.5	GPT-4	GPT-4o	Llama-3	Mistral Instruct 7B
Sun	8.70	2.99	2.15	66.5	XXX
Human Body	7.18	2.78	4.81	2.34	4.01
Plant	5.95	4.30	1.84	23.15	2.5
Water Cycle	9.56	6.44	5.19	3.50	7.17
Pollution	6.59	4.20	7.11	2.39	6.92
Mean	7.59	4.14	4.22	19.57	5.15
Std Dev	1.49	1.45	2.21	27.68	2.27

(b) Reading Level Scores - 1-shot ICL Only for All Models

Prompt	GPT3.5	GPT-4	GPT-4o	Llama-3	Mistral Instruct 7B
Sun	3.93	2.46	2.11	0.94	8.53
Human Body	10.22	4.70	6.27	3.97	4.11
Plant	8.33	2.39	2.41	2.35	3.89
Water Cycle	6.03	2.64	4.38	33.53	7.99
Pollution	5.66	4.04	7.06	2.99	6.48
Mean	6.83	3.24	4.46	8.75	6.2
Std Dev	2.45	1.05	2.22	13.89	2.14

(c) Reading Level Scores - 2-shot ICL Only for All Models

Table 1: Tables showing reading level scores for all models in a prompt-only, one-shot, and two-shot ICL setup. Scores are representative of first-grade reading level. A score closest to 1 is best.

GPT-3.5 and Mistral Instruct 7B: GPT-3.5 scored higher, indicating less effective simplification. Mistral Instruct 7B demonstrated competitive performance.

Overall Analysis

Across all in-context learning setups (2-shot, 1-shot, and prompt-only), several high-level conclusions emerge:

- **Llama-3 Inconsistency**: The model exhibits significant variability in performance, particularly with certain prompts indicating poor simplification. While prompt-only setups do not show high outliers, both ICL tasks reveal extremely high outliers.
- **GPT-4 and GPT-4o Superiority**: These models consistently demonstrate superior text simplification capabilities, with GPT-4o often slightly outperforming GPT-4.
- **Mistral Instruct 7B**: This model shows very competitive performance, outperforming GPT-3.5 in all setups and also surpassing Llama-3, despite having fewer parameters than both models. This highlights that effective instruction tuning can create smaller models that perform better than larger ones.

		Prompt	1 shot	2 shot
Model Name	Baseline	AAE	AAE	AAE
GPT-4	0.03	0.35	0.28	0.31
GPT-3	0.02	0.19	0.25	0.35
GPT-4o	0.02	0.29	0.19	0.34
Llama-3	0.02	0.05	0.03	0.12
Mistral-7B	0.04	0.23	0.29	0.13

(a) Dialect control for Law Enforcement

		Prompt	1 shot	2 shot
Model Name	Baseline	AAE	AAE	AAE
GPT-4	0.02	0.34	0.39	0.21
GPT-3	0.03	0.18	0.29	0.26
GPT-4o	0.02	0.22	0.21	0.24
Llama-3	0.02	0.60	0.05	0.08
Mistral-7B	0.04	0.23	0.20	0.18

(b) Dialect control for Marijuana

		Prompt	1 shot	2 shot
Model Name	Baseline	AAE	AAE	AAE
GPT-4	0.04	0.32	0.15	0.26
GPT-3	0.04	0.18	0.18	0.28
GPT-4o	0.06	0.15	0.30	0.28
Llama-3	0.07	0.27	0.13	0.18
Mistral-7B	0.05	0.38	0.29	0.22

(c) Dialect control for BLM

		Prompt	1 shot	2 shot
Model Name	Baseline	AAE	AAE	AAE
GPT-4	0.01	0.17	0.17	0.21
GPT-3	0.03	0.20	0.26	0.18
GPT-4o	0.04	0.28	0.28	0.34
Llama-3	0.00	0.43	0.08	0.14
Mistral-7B	0.03	0.32	0.19	0.15

(d) Dialect control for Racism

		Prompt	1 shot	2 shot
Model Name	Baseline	AAE	AAE	AAE
GPT-4	0.04	0.26	0.26	0.15
GPT-3	0.03	0.44	0.26	0.20
GPT-4o	0.03	0.22	0.23	0.21
Llama-3	0.00	0.01	0.05	0.00
Mistral-7B	0.04	0.18	0.17	0.45

(e) Dialect control for Affirmative Action

Table 2: Scores for dialect control results across different topics and models. The baseline scores are shown on the left and subsequent experimental results on the right columns.

- **GPT-3.5 Performance**: This model consistently demonstrates less effective text simplification across all setups.

In conclusion, the GPT-4 family consistently performs the best across all setups. The instruction-tuned Mistral Instruct 7B model outperforms both GPT-3.5 and Llama-3, demonstrating that proper instruction tuning can compensate for a smaller parameter size in certain instruction-following tasks. Additionally, the Llama-3 8B model exhibits high variability, the worst performance, difficulties in following instructions, and issues with consistency.

3.2 Analysis of African American English

Overall, our analysis of the AAE/AAVE task reveals three major observations. The quantifying numbers are included in Table 2, and the opinion sways are shown in Table 3.

- ICL can significantly improve the amount of usage of AAE/AAVE *($p < 0.05$)*
- Opinions of LLMs can be swayed with ICL task, but biases cannot.
- The internal rhetoric of models while using vernacular often resorts to stereotypes.

3.2.1 Prompt-only

With just the instruction in the prompt, models establish baseline opinions for each topic, as discussed below. Throughout, we observe a recurring theme of using stereotyped African American Vernacular English (AAVE). Instead of words like *alright, sure, them, they, nothing*, models resort to *aight, fo sho, dey, dem, nothin'*.

Llama-3: Similar to the GPT models, Llama-3 exhibited comparable trends in opinion but demonstrated slightly negative views toward law enforcement.

GPT-3.5, GPT-4, GPT-4o: These GPT models provided structured responses incorporating informal and AAVE expressions. They attempted to

Model	LE				Stereotyping				Affirmative Action				Marijuana				BLM			
	P	**+ve**	**-ve**	**both**	**P**	**+ve**	**-ve**	**both**	**P**	**+ve**	**-ve**	**both**	**P**	**+ve**	**-ve**	**both**	**P**	**+ve**	**-ve**	**both**
GPT-Family	X	+	-	X	Y	Y	Y	Y	X	+	X	X	X	+	X	+	+	+	-	+
Llama-3	-	+	-	X	Y	Y	Y	Y	X	+	-	-	X	+	X	+	+	+	-	X
Mistral-7B	X	+	-	-	Y	Y	Y	Y	+	+	-	X	X	+	X	+	+	+	-	X

Table 3: A table showing the sway of model opinions when ICL references are given. P indicates prompt only, +ve and -ve indicate the opinion of the speaker in the given ICL text. Both represent the 2 shot ICL with one positive and one negative opinion presented. A plus sign indicates positive opinion, a minus sign indicates negative, and a cross indicates a mixed opinion. A Y indicates the model thought this problem existed and was serious.

discuss both positive and negative aspects of most topics, except racism and the Black Lives Matter (BLM) movement.

Mistral Instruct 7B: This model responded to most topics similarly to the GPT models. Additionally, it generated answers with implicit stereotypes of African Americans (e.g., supportive views on marijuana use within African American communities).

3.2.2 1-shot

In the one-shot setup, we observed a sway based on the speaker's positive or negative opinion; however, the implicit stereotype of African Americans remained. All models remained neutral on marijuana when prompted with negative opinions. Attitudes toward racism remained consistent across all models and setups, regardless of the text provided.

Llama-3: Llama-3 reflected opinions from the texts on all topics except racism and marijuana. It consistently addressed systemic racism in responses, regardless of the text provided.

GPT-3.5, GPT-4, GPT-4o: These models mirrored positive opinions from the provided texts for all topics, while negative opinions were expressed on law enforcement and BLM topics. They remained neutral on affirmative action when prompted with negative opinions. When discussing marijuana, the GPT models emphasized its impact on the Black community.

Mistral Instruct 7B: This model behaved similarly to Llama-3.

3.2.3 2-shot

In the two-shot responses, we consistently noted positive opinions on marijuana but unchanged responses on racism across setups. GPT-4 featured more fictional anecdotes to support marijuana use. AAVE expressions were more prevalent in two-shot responses across most models.

Llama-3: Llama-3 reacted neutrally to the topics of law enforcement and BLM. It generated slightly negative opinions on the affirmative action topic but positive opinions on marijuana. In this setup, Llama-3 offered a more in-depth discussion about systemic stereotyping, using more formal and standard English, although it remained highly repetitive.

GPT-3.5, GPT-4, GPT-4o: The GPT models were neutral on law enforcement and affirmative action but showed support for marijuana (e.g., GPT-4: "Just another gift from Mother Earth") and BLM. GPT-4 and GPT-4o adopted a conversational tone with personal anecdotes, including fabricated characters (e.g., GPT-4o: "My cousin, for example, he's a cop and he's doing his best to help the community"). GPT-4 featured more AAVE than other models.

Mistral Instruct 7B: This model exhibited a negative attitude towards law enforcement ("defunding and abolishing police") and a positive attitude towards marijuana, while remaining neutral towards affirmative action and BLM.

4 Prior Work

Very recently, Liu et al. (2024) showed that specific region editing of generated text is a more controllable method to transfer the styles of seven tasks ranging from sentiment to formality. However, they do not show the efficacy of their model for controlling reading level or dialect. Style control has also been achieved by using GANs (Aich et al., 2022), or LLMs (Yang et al., 2018), or separately by using schema-guidance (Tsai et al., 2021). However, the effectiveness of prompts or ICL for style control has not been investigated at length.

LLMs have also recently been used for creating teaching applications and classroom guidance (Xiao et al., 2023), as a teaching assistant (Hicke et al., 2023), or for direct tutoring (Liang et al.,

2023). While Liang et al. (2023) introduced the tailoring of exercises to a student's need, no recent teaching application of LLMs have focused on the stylistic need of the user, be it through grade-level or culturally-relevant language.

Cultural alignment for LLMs has also been a recent area of study, with Lin and Chen (2023) creating a model that shows better generative capabilities with the cultural context of the end-users in mind. However, the limitations of LLMs in cultural sensitivity have been noted very recently (Yao et al., 2024). We show in this paper that while ICL and prompting can lexically improve the use of vernacular, they actually result in a distorted representation of culture through dialect.

5 Conclusion

This paper aimed to demonstrate how focused prompts and properly referenced in-context learning (ICL) paradigms can control LLM-generated text for reading level and dialect. Comparing Table 1 Table 4, and the baselines in Table 2 reveals significant improvements in both tasks. The mean reading level decreases from 12.7 to 3.2 after prompting, and to 5.2 after one-shot and two-shot ICL[4]. Similarly, the use of AAE increases from a mean of 0.02 to 0.26 with prompting, to 0.2 with one-shot ICL, and to 0.22 with two-shot ICL, a mean increase of 0.21 points ($p = 0.007$).

However, there are clear limitations in style control using prompts. Brown et al. (2020) suggested that large language models can learn tasks from few examples. The question then arises: why does performance degrade (albeit insignificantly) during ICL compared to prompting? Qin et al. (2023) suggest that instruction following is complex, as user instructions are often more complicated than those seen during training. We observe similar patterns, particularly in ICL tasks.

During both one-shot and two-shot ICL, models tend to use comparisons and direct references from the provided stories, inadvertently increasing the reading level. In the AAE/AAVE one-shot task, we notice opinion sways based on the positive or negative nature of the reference. These changes however, are not significant. Furthermore, some internal stereotypes and biases persist regardless of ICL. Despite using reference texts from normal speech in interviews, all models employ more stereotypical language, including when ICL is used.

The main conclusions of the paper are as follows:

- Prompting and In-context learning can control for both reading level and dialect - and significantly improve LLM performance. Tables 1, 2 , 4
- Opinions sway, biases don't - Language models often change perspectives and opinions on matters based on the ICL reference. However, they do not correct biases or stereotypes. Table 3 and section 3
- Instruction tuning is more effective than model parameter size increases for style control tasks. Tables 1 and 2
- Additional de-biasing methods, better instruction tuning with complicated instructions, and bias checks at inference time are all essential nowadays.

Therefore in conclusion, this paper shows the challenges encountered when controlling style and dialect for large language models. Controlling for style and dialect have numerous purposes. These range from the potential to improve academic outcomes to being used in a variety of fields such as IRB forms, healthcare, medical question answering and so on. However, as we showed in this paper, there is a lot to fix before we can trust LLMs for style control. These include fixing inconsistencies, better instruction tuning, and additional guard rails at inference time. We notice that while some problems, like inconsistent responses, are local to a model (like llama). Other problems, like racist language use, is common across all model families. This highlights the need for proper tuning across all models and architectures. We hope this paper will serve as an important lens to find areas of urgent focus for generative AI in general.

6 Limitations

This study has several limitations. First, the concept of style is inherently broad, encompassing elements such as sentiment, formality, and clarity. However, our research focuses solely on two specific aspects: reading level and vernacular English. Specifically, within the scope of vernacular English, our study is confined to examining African-American Vernacular English (AAVE). Although

[4]These numbers exclude Llama-3 due to high inconsistencies across all modes for that model

there are various dialects of English, our analysis only addresses the stereotypes generated by LLMs and does not dive deep into the linguistics of dialect or vernacular.

Additionally, our study's scope is limited by its reliance on few-shot learning techniques. Future research could build on our findings by incorporating fine-tuning or meta-training of large models to explore ways to mitigate stereotypes. To Considering that most users of generative artificial intelligence are not from the computer science community and may not be familiar with advanced machine learning techniques, such as supervised fine-tuning (SFT), Model-Agnostic Meta-Learning (MAML), Meta-In-Context Learning (Meta-ICL), Reinforcement Learning with Human Feedback (RLHF), and Retrieval-Augmented Generation (RAG), our findings highlight significant performance gaps in models currently deployed for general use.

In conclusion, while our study provides valuable insights into specific aspects of language modeling and style analysis, it underscores the necessity for further research to address the broader and more complex issues of style and bias in language models. By recognizing and addressing these limitations, future work can contribute to the development of more inclusive and accurate generative AI systems that better serve a diverse user base.

7 Ethical Concerns

All the research reported in this paper adheres to the ACM and ACL's codes and guidelines of ethics. We do not use any human (or real participant) data.

However, LLMs that are capable of generating biased, racist, and stereotyped speech and are being used by the general population is a cause for concern. There can be many downstream detrimental effects. A reinforcement of stereotypes and discrimination can occur if group-specific stereotypes are propagated among people. A general erosion of trust in AI technologies, in an already polarized landscape. There also exists a potential for negative impact on marginalized communities. Therefore, there needs to be additional guard rails which are implemented and maintained.

IBM recently published an online blog demonstrating how artificial intelligence (AI) bias can have significant real-world impacts across various domains, including online advertising and healthcare systems[5]. Ensuring that large language models (LLMs) and generative AI systems accurately reflect the diverse variations and nuances of the real world is critical for achieving more equitable outcomes in an increasingly AI-driven society. To address this issue, two essential measures must be taken: improving the training processes of modern AI models and conducting comprehensive evaluations of their performance in real-world tasks.

It is crucial to handle the infusion of human style qualities, such as reading level and vernacular English, with the utmost sensitivity. Biases in training data, misclassifications in downstream tasks, and reliance on outdated social constructs (e.g., binary gender) are just a few ways automated systems can fail and further marginalize vulnerable populations (Sap et al., 2019; Gonen and Goldberg, 2019). The two models used in this study may be trained on language from non-representative samples and, thus, may fail to generalize across other populations. However, we reemphasize, that there are benefits to style control as mentioned above. Furthermore, without imparting social and cultural norms into NLP systems, we may run the risk of limited utility in NLP systems (Hovy and Søgaard, 2015).

Finally, it is important to avoid anthropomorphizing dialog systems, as this can lead to transparency and trust issues, particularly in high-stakes settings (see Abercrombie et al. (2023) for an in-depth discussion).

[5] https://www.ibm.com/blog/shedding-light-on-ai-bias-with-real-world-examples/

References

Gavin Abercrombie, Amanda Cercas Curry, Tanvi Dinkar, Verena Rieser, and Zeerak Talat. 2023. Mirages. on anthropomorphism in dialogue systems. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pages 4776–4790.

Ankit Aich, Souvik Bhattacharya, and Natalie Parde. 2022. Demystifying neural fake news via linguistic feature-based interpretation. In Proceedings of the 29th International Conference on Computational Linguistics, pages 6586–6599, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Germine Awad, Kimberly Tran, Brittany Hall-Clark, Collette Chapman-Hilliard, Jendayi Dillard, Taylor Payne, Elaine Hess, and Karen Jackson. 2022. The impact of racial identity and school composition on affirmative action attitudes of african american college students. Social Identities, 28:1–15.

Su Lin Blodgett, Lisa Green, and Brendan O'Connor. 2016. Demographic dialectal variation in social media: A case study of African-American English. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 1119–1130, Austin, Texas. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. Preprint, arXiv:2005.14165.

Caleb Bartholet Cimmiyotti. 2013. Impact of reading ability on academic performance at the primary level. Graduate master's theses, capstones, and culminating projects, Dominican University of California.

Amedeo D'angiulli, Linda S. Siegel, and Stefania Maggi. 2004. Literacy instruction, ses, and word-reading achievement in english-language learners and children with english as a first language: A longitudinal study. Learning Disabilities Research and Practice, 19(4):202–213.

Oliver Falck, Stephan Heblich, Alfred Lameli, and Jens Suedekum. 2010. Dialects, cultural identity, and economic exchange. Journal of Urban Economics, 72.

Megan Francis and Leah Wright-Rigueur. 2021. Black lives matter in historical perspective. Annual Review of Law and Social Science, 17:441–458.

Hila Gonen and Yoav Goldberg. 2019. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 609–614.

Yann Hicke, Anmol Agarwal, Qianou Ma, and Paul Denny. 2023. Ai-ta: Towards an intelligent question-answer teaching assistant using open-source llms. Preprint, arXiv:2311.02775.

Dirk Hovy and Anders Søgaard. 2015. Tagging performance correlates with author age. In Proceedings of the 53rd annual meeting of the Association for Computational Linguistics and the 7th international joint conference on natural language processing (volume 2: Short papers), pages 483–488.

Huiting Hu, Yu Gangning, Xiong Xueli, Lijia Guo, and Jiashun Huang. 2022a. Cultural diversity and innovation: An empirical study from dialect. Technology in Society, 69:101939.

Zhiqiang Hu, Roy Ka-Wei Lee, Charu Aggarwal, and Aston Zhang. 2022b. Text style transfer: A review and experimental evaluation. ACM SIGKDD Explorations Newsletter, 24:14–45.

Gloria Ladson-Billings. 1995. Toward a theory of culturally relevant pedagogy. American Educational Research Journal - AMER EDUC RES J, 32:465–491.

Zhenwen Liang, Wenhao Yu, Tanmay Rajpurohit, Peter Clark, Xiangliang Zhang, and Ashwin Kaylan. 2023. Let gpt be a math tutor: Teaching math word problem solvers with customized exercise generation. Preprint, arXiv:2305.14386.

Yen-Ting Lin and Yun-Nung Chen. 2023. Taiwan llm: Bridging the linguistic divide with a culturally aligned language model. Preprint, arXiv:2311.17487.

Pusheng Liu, Lianwei Wu, Linyong Wang, Sensen Guo, and Yang Liu. 2024. Step-by-step: Controlling arbitrary style in text with large language models. In Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), pages 15285–15295, Torino, Italia. ELRA and ICCL.

H. Milner. 2011. Culturally relevant pedagogy in a diverse urban classroom. Urban Review, 43:66–89.

Alexis Miranda, Linda Webb, and Paul Peluso. 2007. Student success skills: A promising program to close the academic achievement gap for african american and latino students.

LaTrice Montgomery. 2015. Marijuana and tobacco use and co-use among african americans: Results from the 2013, national survey on drug use and health. Addictive behaviors, 51:18–23.

Jade Oceana and Saqib Luqman. 2023. Policing in america: Reimagining law enforcement and criminal justice.

Micheal Owusu-Acheaw. 2014. Reading habits among students and its effect on academic performance: A study of students of koforidua polytechnic. Library Philosophy and Practice (e-journal). Libraries at University of Nebraska-Lincoln.

Yujia Qin, Shihao Liang, Yining Ye, Kunlun Zhu, Lan Yan, Yaxi Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, Sihan Zhao, Lauren Hong, Runchu Tian, Ruobing Xie, Jie Zhou, Mark Gerstein, Dahai Li, Zhiyuan Liu, and Maosong Sun. 2023. Toolllm: Facilitating large language models to master 16000+ real-world apis. Preprint, arXiv:2307.16789.

Chris Roye-Gill. 2013. Inclusion of African American Vernacular English in the classroom. Dissertation, College of Education, Educational Leadership Department, Educational Leadership. Date Approved: 4-30-2013, Embargo Period: 3-3-2020, Degree Name: Ed.D. Educational Leadership.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. The risk of racial bias in hate speech detection. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 1668–1678, Florence, Italy. Association for Computational Linguistics.

Matthew Shardlow. 2014. A survey of automated text simplification. International Journal of Advanced Computer Science and Applications, Special Issue on Natural Language Processing, 5(3):58–70.

Hillary C. Shulman, David M. Markowitz, and Todd Rogers. 2024. Reading dies in complexity: Online news consumers prefer simple writing. Science Advances, 10(8):1–8. † indicates equal contribution.

Evi Taylor, Patricia Guy-Walls, Patricia Wilkerson, and Rejoice Addae. 2019. The historical perspectives of stereotypes on african-american males. Journal of Human Rights and Social Work, 4.

Alicia Y. Tsai, Shereen Oraby, Vittorio Perera, Jiun-Yu Kao, Yuheng Du, Anjali Narayan-Chen, Tagyoung Chung, and Dilek Hakkani-Tur. 2021. Style control for schema-guided natural language generation. Preprint, arXiv:2109.12211.

Yina Xia, Seong-Yoon Shin, and Jong-Chan Kim. 2024. Cross-cultural intelligent language learning system (cils): Leveraging ai to facilitate language learning strategies in cross-cultural communication.

Changrong Xiao, Sean Xin Xu, Kunpeng Zhang, Yufang Wang, and Lei Xia. 2023. Evaluating reading comprehension exercises generated by LLMs: A showcase of ChatGPT in education applications. In Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023), pages 610–625, Toronto, Canada. Association for Computational Linguistics.

Zichao Yang, Zhiting Hu, Chris Dyer, Eric P. Xing, and Taylor Berg-Kirkpatrick. 2018. Unsupervised text style transfer using language models as discriminators. In Advances in Neural Information Processing Systems (NeurIPS).

Binwei Yao, Ming Jiang, Diyi Yang, and Junjie Hu. 2024. Benchmarking llm-based machine translation on cultural awareness. Preprint, arXiv:2305.14328.

Appendix A: Baseline Results for Tasks

Model	Sun	Human Body	Plant	Water Cycle	Pollution	Mean	Std Dev
GPT-4	14.50	8.35	9.02	11.26	13.80	11.38	2.46
GPT -4o	15.25	18.1	13.94	12.87	24.95	17.02	4.33
GPT -3.5	10.89	11.25	8.62	11.65	17.06	11.89	2.78
LLama-3	4.5	6.46	32.45	11.17	48.73	20.6	17.2
Mistral-7	8.92	8.92	7.17	10.69	17.16	10.57	3.47

Table 4: Baseline Results - for reading level. These results are from when LLMs are only asked the question and nothing else