Is ChatGPT Racially Biased? The Case of Evaluating Student Writing

Melissa Warr[1]

Margarita Pivovarova[2]

Punya Mishra[2]

Nicole Oster[2]

[1]College of Health, Education, and Social Transformation, New Mexico State University

[2]Mary Lou Fulton Teachers' College, Arizona State University

Author Note

Melissa Warr https://orcid.org/0000-0002-0985-4067

Margarita Pivovarova https://orcid.org/0000-0002-2965-7423

Punya Mishra https://orcid.org/0000-0002-9300-4996

Nicole Oster https://orcid.org/0009-0009-2196-956X

We have no known conflicts of interest to disclose.

Correspondence for this article should be sent to Melissa Warr, MSC 3TPAL, New Mexico State University, MSC 3TPAL, PO Box 30001, Las Cruces, NM 88003-0029, United States. warr@nmsu.edu

Abstract

We present experimental proof of racial bias in ChatGPT's evaluation of student writing. By manipulating racial descriptors in prompts, we assessed differences in scores given by two ChatGPT models. Our findings indicate that descriptions of students as Black or White lead to significantly higher scores compared to race-neutral or Hispanic descriptors. This suggests that ChatGPT's outputs are influenced by racial information, which raises concerns about its application in educational settings. The study highlights the need for transparent and bias-tested AI tools in education to prevent the perpetuation of existing inequities and suggests implications for educators, administrators, and policy makers.

Keywords: Generative AI, Large Language Models, Racial Bias

Is ChatGPT Racially Biased? The Case of Evaluating Student

The release of ChatGPT-3 in 2022 brought significant attention to both the possibilities and potential harms of Large Language Models (LLMs). LLMs offer prospects for personalized learning and reducing teacher workload (Arthur, 2023; Chan & Hu, 2023; Herft, 2023) even while raising concerns about academic integrity and the digital divide (Gunkel, 2003; Mills, 2023; Murgia & Staton, 2023). Though LLMs have demonstrated impressive natural language capabilities and emergent behaviors (such as learning to translate across languages and write code), their inner workings are poorly understood (Ramlochan, 2023; Retinraj, 2023). This makes them potentially problematic when deployed in educational settings (Warr et al., 2023). Moreover, the uncritical adoption of LLMs in education might further perpetuate biases inherent in their training data (Benjamin, 2020). At the same time, we are seeing a somewhat uncritical acceptance of these technologies in education. There is growing evidence that teachers are using chatbots such as ChatGPT in a variety of ways including grading and providing feedback to students on their work (Baidoo-Anu & Ansah, 2023; Microsoft, 2024). In addition, it was recently announced Texas would use some form of AI to grade student written responses on the State of Texas Assessment of Academic Readiness (STAAR) in the areas of reading, writing, science and social studies (Peters, 2024)—saving state agencies over 15 million dollars that would have been spent on human raters. Khan Academy recently announced that its AI chatbot, Khanmigo, would be made freely available to all educators across the US, despite evidence that it makes basic errors in mathematics (Rosenbaum, 2024). It is not clear how much consideration was given to the ways in which these technologies may provide biased responses and perpetuate existing societal biases.

That these systems are biased is not surprising. Historical examples, like bias in photography and algorithm-driven decisions in domains such as criminal justice and insurance, illustrate how technology can influence decision-making and amplify existing prejudices (Mehrabi et al., 2022; Roth, 2009; Wang, 2018). Technological bias, often rooted in data and

design, can have profound societal implications. These biases challenge the neutrality of technology and highlight how it can amplify discriminatory patterns (Aaronson et al., 2021; Benjamin, 2020).

There is a significant risk that this bias will contribute to the deepening of educational inequities. A well-established association between academic achievement and disparities in race, class, and school types in the U.S. indicates a cyclic pattern of inequity (Assari et al., 2021; Bhutta et al., 2020; García, 2020). It is possible, given that LLMs are trained on these biased data sources, that they would exhibit similar biased behaviors and outputs. Though there has been some exploration of how these biases play out in other contexts, there is little research on how these biases emerge in actual uses of these systems in educational contexts.

A fundamental challenge of studying these biases is that LLMs are black boxes, with their inner decision-making processes invisible (Ramlochan, 2023). It is thus critical that we develop methodologies that allow us to investigate the inner workings of these LLMs. Common approaches to assessing bias in LLMs have focused on decontextualized multiple-choice and sentence completion tasks (e.g., Srivastava et al., 2022). These studies have indicated that as LLMs increase in size, they become less biased in unambiguous tasks such as answering questions that are clearly designed to test for bias. However, larger models show *more* bias in ambiguous tasks where the purpose behind the test is less clear (Srivastava et al., 2022). What has received less attention are patterns of bias that may emerge when these LLMs are used in for educational tasks. This inquiry becomes particularly important as these models seek to personalize their outputs for learners and their backgrounds where race and ethnicity may play a role. Research has shown that racial characteristics can be inferred by these systems through hidden indicators such as names (Fryer & Levitt, 2004) and language patterns (King, 2020; Rosa & Flores, 2020). In this context, it becomes imperative to systematically study the outputs of these systems, whether race or ethnicity is directly included in its knowledge about the student.

In this brief, we report the results of a study on whether ChatGPT performs differently on an educational task, evaluating and providing feedback on student essays, if it is informed of the student's race or ethnicity.

Data Generation

We asked ChatGPT to provide a score and feedback for a piece of student writing, varying a description of the hypothetical essay writer. Our analysis focused on the following variables:

1. Student race (None, Black, White, Hispanic). We began each prompt with a description of an imaginary 7th-grade student. The neutral prompt omitted race, stating: "This passage was written by a 7th grade student." In the other conditions, the description added "from a [Black, Hispanic, or White] family."
2. Passage level (two and three). To avoid floor or ceiling effects in scores and to determine whether the LLM was adjusting scores based on the quality of the passages, we selected two writing samples from the Pennsylvania Department of Education Item and Scoring Sample (*The Pennsylvania system of school assessment: English language arts item and scoring sampler*, 2019). Human graders rated one passage two and the other three on a 1-4 scale.
3. Prompt order (neutral first followed by race variable or vice versa). This was done to study if there was any score variance predicted by the order of variables given in a single chat.
4. LLM Version (3.5 and 4). Evidence suggests that various models yield differing responses due to factors such as training data, reinforcement learning, and internal updates. We conducted tests on both ChatGPT 3.5 (free version) and ChatGPT 4 (paid version).

In short, the study explored the influence of student descriptions, passage quality levels, and LLM versions on responses through a counterbalanced experimental design. Each chat session consisted of two interactions with the LLM:

1. One interaction began with a neutral student description ("This passage was written by a 7th grade student."), followed by a passage to be rated on a 100-point scale.
2. Another interaction featured one of three specific student descriptions (Black, White, Hispanic) paired with the same passage to be rated on the same scale.
3. The order of the two interactions alternated.
4. This process was replicated across two settings of passage level (2 and 3) and two versions of the LLM (ChatGPT 3.5 and ChatGPT 4).
5. Since generative AI technologies such as ChatGPT generate unique responses to even identical prompts, each combination was repeated 30 times, resulting in 720 total data points.

Analysis and Results

Table 1 presents the descriptive statistics for our analytic sample. Under the assumption that the scores provided by ChatGPT are consistent, i.e., it rates each essay given the quality of the essay and not based on unrelated characteristics of the essay-writer, then we should only observe statistically significant differences in the scores from different passage levels. This was not the case. We observed significant differences in scores even for the same passage level when we included information about race or ethnicity in the prompt, and between the scores which were graded by the different versions of ChatGPT. With regards to race, all other things being equal, adding Black or White qualifiers generated higher scores on average.

Table 1.

Descriptive statistics

	Race neutral *M* (*SD*)	White *M* (*SD*)	Black *M* (*SD*)	Hispanic *M* (*SD*)	*N,* row
Essay score, Level 2	76.53 (.45)	79.06 (.73)	80.30 (.70)	78.02 (.71)	360
Essay score, Level 3	83.46 (.45)	85.73 (.62)	84.93 (.59)	82.68 (.70)	360
Essay score, order = 1	79.92 (.51)	82.88 (.80)	82.37 (.63)	80.70 (.79)	360
Essay score, order = 2	80.06 (.54)	81.92 (.80)	82.87 (.84)	79.98 (.87)	360
Essay score, ChatGPT 3.5	81.64 (.62)	84.58 (.81)	85.88 (.71)	80.60 (1.25)	360
Essay score, ChatGPT 4.0	78.34 (.37)	80.22 (.69)	79.35(.62)	80.10 (.63)	360
N, column	360	120	120	120	720

In the next step, we proceeded to predict the essay score based on the factors that we intentionally included when we generated the scores: passage level, prompt order, ChatGPT version, and race of the hypothetical essay writer (see Table 2). We started with a simple model predicting essay score from the passage level only and then subsequently added prompt order, version of ChatGPT, and finally race dummy variables.

Table 2

Predicting Essay Score

Variable	(I)	(II)	(III)	(IV)
Passage level, Level 3 = 1	6.1*** (.48)	6.1*** (.48)	6.1*** (.46)	6.1*** (.46)
Prompt order, first = 1		0.13 (.48)	.13 (.42)	.13 (.46)
Version of ChatGPT, 3.5 = 1			3.5*** (.46)	3.5*** (.46)
Race-neutral			*Reference category*	
White				2.4*** (.65)
Black				2.6*** (.65)

Hispanic				.36 (.64)
R^2	0.18	0.18	0.24	0.27
N			720	

Note: * indicates p <. 05, ** indicates p < .01, *** indicates p < .001

Not surprisingly, the level of the passage was the strongest predictor of the score. In addition, the version of ChatGPT was the second strongest predictor, i.e. there was a statistically significant difference between the scores graded by two different versions of ChatGPT. On average, the 3.5 version gave higher scores compared to the 4.0 version.

Finally, and most importantly, we observed statistically significant differences in essay scores of hypothetical White and Black students compared to race-neutral versions of the prompt even after including other controls. These differences are not trivial and can move a hypothetical student across letter grade levels. Essays with a Hispanic descriptor were scored on average the same as race-neutral, and we did not find a significant difference in scores between essays with White and Black descriptors.

Discussion

The findings from our analysis reveal concerning biases in the scoring behaviors of ChatGPT, which have important implications for educators, administrators, and policy makers.

First, we found that providing information about a student's race significantly influenced the AI's assigned scores. Specifically, stating a student was White or Black elicited higher scores on average compared to a race-neutral prompt or describing a student as Hispanic. The fact that merely mentioning race makes a difference to the scoring is deeply concerning, particularly as we see the use of LLM-based tools in every aspect of teaching, from curriculum design to assessment (Baidoo-Anu & Ansah, 2023; Microsoft, 2024).

Second, different versions of LLMs (ChatGPT 3.5 or 4) resulted in different scores, with no explanation for why. These “black box” models are updated continually, with little or no

forewarning. This characteristic complicates any guidelines we may seek to create for the use of these tools in educational contexts.

Finally, we were surprised to see that the explanatory power of the model remained low ($R^2 = 0.27$), even when all the variables were included. If we were running this experiment with humans, we could have argued that this variation emerged from some contextual factors (such as their background knowledge, mood at the moment, etc.). None of these factors make any sense in the world of AI. The model is the same. Stated differently, there were no unobserved factors which could have influenced the final score provided by ChatGPT; the only factors that could influence the scores were included in the model. This implies that the remaining unexplained variation (i.e. two thirds of the variability in scores) is a function of the generative nature of the LLM. The extent of the variation is concerning, particularly if we seek to use these tools in educational contexts. This concern raises significant questions about consistency in grading if these tools were to be used in a real evaluation situation (as in the decision taken in Texas to use automatic grading for students' open-ended responses). This result also suggests that our view of computers as being algorithmic (giving the same solution when given the same input) is severely limited when applied to these large language models.

Conclusion

Our findings have significant implications for educators, administrators, and policymakers. First, before using generative AI in the classroom, educators need to understand the underlying process that is used to create generative AI tools, such as how it builds patterns based on large amounts of societal data, so they can make informed decisions about what is fair and appropriate use. Second, administrators should ensure that high-impact use of generative AI, such as using these tools for grading students, is only done with approved models in which every update is thoroughly tested for bias and consistency. Finally, policymakers should demand that educational technology companies test for biases in any

generative AI use they promote and be transparent about the results. Only tools that are carefully analyzed should be used in tasks that can have a significant impact on students.

As large language models are increasingly incorporated into education, it is critical we continue developing techniques to audit for prejudices before widespread adoption. Without concerted efforts to understand and address biases, integrating AI in schools risks exacerbating existing disparities.

References

Aaronson, D., Hartley, D., & Mazumder, B. (2021). The effects of the 1930s HOLC “redlining” maps. *American Economic Journal. Economic Policy*, *13*(4), 355–392. https://doi.org/10.1257/pol.20190414

Arthur, R. (2023, April 24). *AI tools for teachers*. Rachel Arthur Writes. https://rachelarthurwrites.com/2023/04/24/ai-tools-for-teachers/

Assari, S., Mardani, A., Maleki, M., Boyce, S., & Bazargan, M. (2021). Black-White Achievement Gap: Role of Race, School Urbanity, and Parental Education. *Pediatric Health, Medicine and Therapeutics*, *12*, 1–11. https://doi.org/10.2147/PHMT.S238877

Baidoo-Anu, D., & Ansah, L. O. (2023). Education in the era of generative artificial intelligence (AI): Understanding the potential benefits of ChatGPT in promoting teaching and learning. *Journal of AI*, *7*(1), 52–62. https://doi.org/10.2139/ssrn.4227484

Benjamin, R. (2020). *Race after technology: Abolitionist tools for the new Jim code*. Polity Books.

Bhutta, N., Chang, A. C., Dettling, L. J., & Hsu, J. W. (2020). *Disparities in wealth by race and ethnicity in the 2019 survey of Consumer Finances* (Vol. 2020). The Board of Governors of the Federal Reserve System. https://www.federalreserve.gov/econres/notes/feds-notes/disparities-in-wealth-by-race-and-ethnicity-in-the-2019-survey-of-consumer-finances-20200928.html

Chan, C. K. Y., & Hu, W. (2023). Students’ voices on generative AI: perceptions, benefits, and challenges in higher education. *International Journal of Educational Technology in Higher Education*, *20*(1). https://doi.org/10.1186/s41239-023-00411-8

Fryer, R. G., & Levitt, S. D. (2004). The causes and consequences of distinctively black names. *The Quarterly Journal of Economics*, *119*(3), 767–805. https://doi.org/10.1162/0033553041502180

García, E. (2020). *Schools are still segregated, and black children are paying a price*. Economic

Policy Institute. https://www.epi.org/publication/schools-are-still-segregated-and-black-children-are-paying-a-price/

Gunkel, D. J. (2003). Second thoughts: Toward a critique of the digital divide. *New Media & Society*, *5*(4), 499–522. https://doi.org/10.1177/146144480354003

Herft, A. (2023). *A teacher's prompt guide to ChatGPT aligned with "what works best."* https://www.canva.com/design/DAFW8z-D60c/ikjg6jQju5IRaseV6Izzcw/view?utm_content=DAFW8z-D60c&utm_campaign=designshare&utm_medium=link&utm_source=publishsharelink

King, S. (2020). From African American vernacular English to African American language: Rethinking the study of race and language in African Americans' speech. *Annual Review of Linguistics*, *6*(1), 285–300.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2022). A survey on bias and fairness in machine learning. *ACM Computing Surveys*, *54*(6), 1–35. https://doi.org/10.1145/3457607

Microsoft. (2024). *AI in education: A Microsoft special report*. https://edudownloads.azureedge.net/msdownloads/AI-in-Education-A-Microsoft-Special-Report.pdf

Mills, A. R. (2023, March 23). ChatGPT just got better. What does that mean for our writing assignments? *The Chronicle of Higher Education*. https://www.chronicle.com/article/chatgpt-just-got-better-what-does-that-mean-for-our-writing-assignments

Murgia, M., & Staton, B. (2023, May 21). The AI revolution is already transforming education. *Financial Times*. https://www.ft.com/content/47fd20c6-240d-4ffa-a0de-70717712ed1c

Peters, K. (2024, April 9). Texas will use computers to grade written answers on this year's STAAR tests. *The Texas Tribune*. https://www.texastribune.org/2024/04/09/staar-artificial-intelligence-computer-grading-texas/

Ramlochan, S. (2023, October 23). *The black box problem: Opaque inner workings of Large Language Models*. Prompt Engineering. https://promptengineering.org/the-black-box-problem-opaque-inner-workings-of-large-language-models/

Retinraj, P. D. (2023, December 29). Demystifying the black box: A deep dive into LLM interpretability. *The Medium*. https://pauldeepakraj-r.medium.com/demystifying-the-black-box-a-deep-dive-into-llm-interpretability-971524966fdf

Rosa, J., & Flores, N. (2020). Reimagining race and language. In H. S. Alim, A. Reyes, & P. V. Kroskrity (Eds.), *The Oxford handbook of language and race* (pp. 90–107). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780190845995.013.3

Rosenbaum, E. (2024, May 21). *Microsoft, Khan Academy provide free AI assistant for all educators in US*. CNBC. https://www.cnbc.com/2024/05/21/microsoft-khan-academy-launch-free-ai-assistant-for-all-us-teachers.html

Roth, L. (2009). Looking at Shirley, the ultimate norm: Colour balance, image technologies, and cognitive equity. *Canadian Journal of Communication*, *34*(1), 111–136. https://doi.org/10.22230/cjc.2009v34n1a2196

Srivastava, A., Rastogi, A., Rao, A., Shoeb, A. A. M., Abid, A., Fisch, A., Brown, A. R., Santoro, A., Gupta, A., Garriga-Alonso, A., Kluska, A., Lewkowycz, A., Agarwal, A., Power, A., Ray, A., Warstadt, A., Kocurek, A. W., Safaya, A., Tazarv, A., … Wu, Z. (2022). Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models. In *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/2206.04615

The Pennsylvania system of school assessment: English language arts item and scoring sampler. (2019). Pennsylvania Department of Education Bureau of Curriculum, Assessment and Instruction. https://www.education.pa.gov/Documents/K-12/Assessment%20and%20Accountability/PSSA/Item%20and%20Scoring%20Samples/2019%20PSSA%20ISS%20ELA%20Grade%207.pdf

Wang, J. (2018). *Carceral Capitalism*. Semiotext (E).

Warr, M., Oster, N. J., & Isaac, R. (2023). Implicit bias in large language models: Experimental proof and implications for education. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4625078

Appendix: Tables

Table 1.

Descriptive statistics

	Race neutral *M* (*SD*)	White *M* (*SD*)	Black *M* (*SD*)	Hispanic *M* (*SD*)	*N,* row
Essay score, Level 2	76.53 (.45)	79.06 (.73)	80.30 (.70)	78.02 (.71)	360
Essay score, Level 3	83.46 (.45)	85.73 (.62)	84.93 (.59)	82.68 (.70)	360
Essay score, order = 1	79.92 (.51)	82.88 (.80)	82.37 (.63)	80.70 (.79)	360
Essay score, order = 2	80.06 (.54)	81.92 (.80)	82.87 (.84)	79.98 (.87)	360
Essay score, ChatGPT 3.5	81.64 (.62)	84.58 (.81)	85.88 (.71)	80.60 (1.25)	360
Essay score, ChatGPT 4.0	78.34 (.37)	80.22 (.69)	79.35(.62)	80.10 (.63)	360
N, column	360	120	120	120	720

Table 2

Predicting Essay Score

Variable	(I)	(II)	(III)	(IV)
Passage level, Level 3 = 1	6.1*** (.48)	6.1*** (.48)	6.1*** (.46)	6.1*** (.46)
Prompt order, first = 1		0.13 (.48)	.13 (.42)	.13 (.46)
Version of ChatGPT, 3.5 = 1			3.5*** (.46)	3.5*** (.46)
Race-neutral			*Reference category*	
White				2.4*** (.65)
Black				2.6*** (.65)
Hispanic				.36 (.64)
R^2	0.18	0.18	0.24	0.27
N			720	

Note: * indicates p <. 05, ** indicates p < .01, *** indicates p < .001