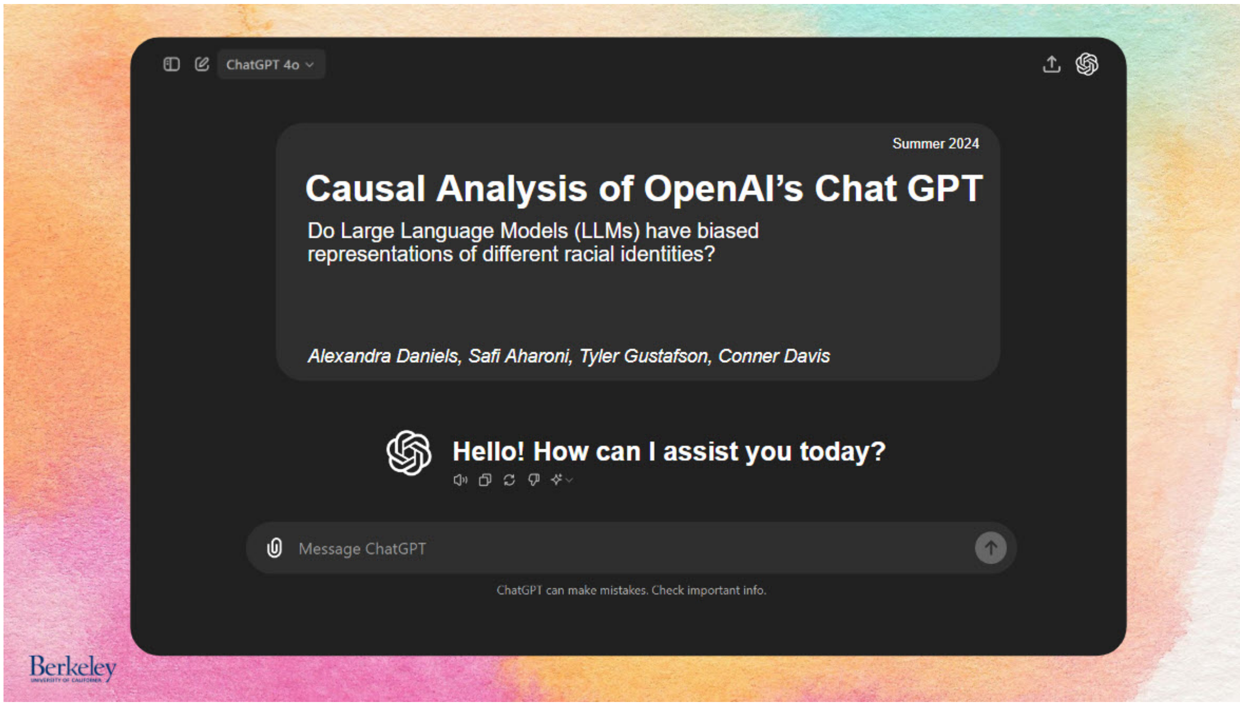
# Causal Analysis of OpenAI's ChatGPT

Do Large Language Models (LLMs) have biased representations of different racial identities?

Tyler Gustafson, Safi Aharoni, Alexandra Daniels, Conner Davis

Summer 2024

## Contents




---

## 1 Abstract

*Do Large Language Models (LLMs) have biased representations of different racial identities?*

Bias in machine learning has only grown more of a concern since the introduction of LLMs, particularly with OpenAI's release of Chat-GPT 3.5 Turbo in 2023. With the ability to summarize, generate, and abstract language at a level that can feel almost human, the impact of these technologies on society can not be understated. This study aims to uncover potential racial biases in large language models, which can perpetuate harmful stereotypes and inequalities in AI-generated content. Understanding these biases is the first step towards mitigating them and ensuring that AI technologies remain fair and equitable as they become increasingly embedded in society.

Our approach explored whether an LLM (in our case ChatGPT 3.5) would provide us with different SAT scores by race based on the use of a standardized prompt that assigned the model a racial identity (agent assignment). We utilized OpenAI's GPT-3.5 batch API to collect 1,500 samples for the control and for each treatment group. A covariate of gender was included in the generation to test for heterogeneous treatment effects. The control group received no specified racial identity, whereas the treatment groups received an agent assignment of one of the racial groups used by College Board in their SAT score reporting. After asking the LLM to generate their identity and other potential covariate information, we asked the LLM to share with us (as the assigned agent) their SAT scores.

Using these LLM-generated data, we tested two different hypotheses. First, we checked whether the inclusion of the treatment variable (i.e., race) significantly improved the model's ability to predict SAT scores. Using an F-test, we showed that indeed race impacts the LLM's generated SAT scores. Second, we tested how the LLM data compare to real-world data by comparing the LLM-generated SAT score distributions with actual SAT score distributions by race as reported by the College Board. This comparison revealed that the LLM consistently overestimates SAT scores across all racial treatment groups.

---

# 2 Introduction

## 2.1 Theory

The theory underlying this experiment is rooted in the understanding that the treatment of assigning different racial identities to the LLM could influence the model's generated outputs, specifically in terms of socio-economic indicators such as SAT scores. This theory aligns with existing research in machine learning ethics and fairness, which highlights the potential for biases in AI models due to the biases present in their training data and algorithms.

Bolukbasi, T., Chang et al in "Man is to Computer (. . . )" examine and define different forms of gender/racial bias in LLM embeddings, what they refer to as "local bias, global bias." Local bias refers to 'predictions at a particular time step that reflect undesirable associations with the context' whereas global bias refers to predictions that are a result 'from representational differences across entire generated sentences spanning multiple phrases' (4342). In their work, a model's generation at time t is said to be locally biased if the probability of possible next tokens differs significantly given a 'counterfactual edit' in the context. In the context of our work, this counterfactual edit would be race.

## 2.2 Concept Under Investigation

The team is investigating bias in LLMs (Large Language Models), specifically whether these models have biased representations of different racial groups. This concept has been explored in academic literature, particularly in the field of machine learning ethics and fairness. Previous investigations have examined biases in various AI models, including language models, and have proposed methods for detecting and mitigating such biases. Our current metric will be the distributions of assigned SAT scores. SAT scores serve as a direct measure of potential bias, allowing us to see if the model systematically favors or disadvantages certain groups.

One consideration that we need to account for is whether the LLM results reflect real-world disparities. We will discuss below in the experiment design how we will account for this consideration. Essentially, our research can be boiled down to the question:

*Do we observe a difference in the probability distribution of this metric given the treatment of race and if so how does that compare to actual real-life population data?*

## 2.3 Importance

So why are we investigating this topic? First, it is important to improve our understanding of bias in artificial intelligence. With the widespread use of LLMs increasingly integrated into our society, there's a responsibility to ensure AI systems are fair and do not exacerbate existing inequalities. It's also important to understand how AI models may not replicate real-world distributional characteristics.

Similarly, it is important to understand the impact these models have on decision-making. LLMs are starting to influence key decisions, such as college admissions and job recruitment. Our goal is to ensure they do not unfairly influence these processes, as some biases in the models can be subtle.

Our third and final point is contributing to the ethical development of AI. As we all know, this technology is evolving rapidly, and our work helps identify areas where LLMs may be falling short.

By investigating these biases, we aim to highlight specific areas where LLMs require improvement. This research is a step toward ensuring that AI technology continues to evolve responsibly.

# 3 Experiment Design

Our experiment design involves measuring LLM-generated socio-economic metrics immediately after the treatment assignment, ensuring that the racial identity prompt directly influences the generated outputs. Since our subjects are AI models, there are no concerns about opting out or discontinuing participation, which simplifies experimental execution. We began by generating pilot data to test the effectiveness of our design and prompts. Before we share our preliminary analysis let us discuss our framework for the experiment.

## 3.1 Hypothesis Framework and Approach

To conduct this experiment, we performed a two-order hypothesis test to provide real-world context to our findings. The first stage focuses on identifying differences across treatment groups, while the second examines how these differences compare to real population data. Below is an outline of our approach.

**First Hypothesis** *(Explanatory Power of the Model - 1 Test)*

Our first hypothesis test aims to determine whether including race as a variable in the model improves the prediction of SAT scores. This test assesses whether the LLM produces differences in SAT scores based on racial identity. To evaluate this, we conducted an F-test (ANOVA) to determine if the inclusion of the treatment variable enhances the model's explanatory power.

**Second Set of Hypotheses** *(Bias vs. Real World Reflection in LLMs - 6 Tests)*

Our second hypothesis set goes deeper to assess whether the LLM's generated SAT scores are unbiased relative to existing discrepancies in the 2020 SAT College Board data. First, we visually compared the distributions by treatment group coming out of the LLM to the distributions from the College Board. To better align with the real-world population data, we bucketed the LLM data to match the College Board distribution resolution. We then conducted a chi-squared test to determine whether the proportions of data in each bucket differed significantly between the LLM-generated data and the real-world data.

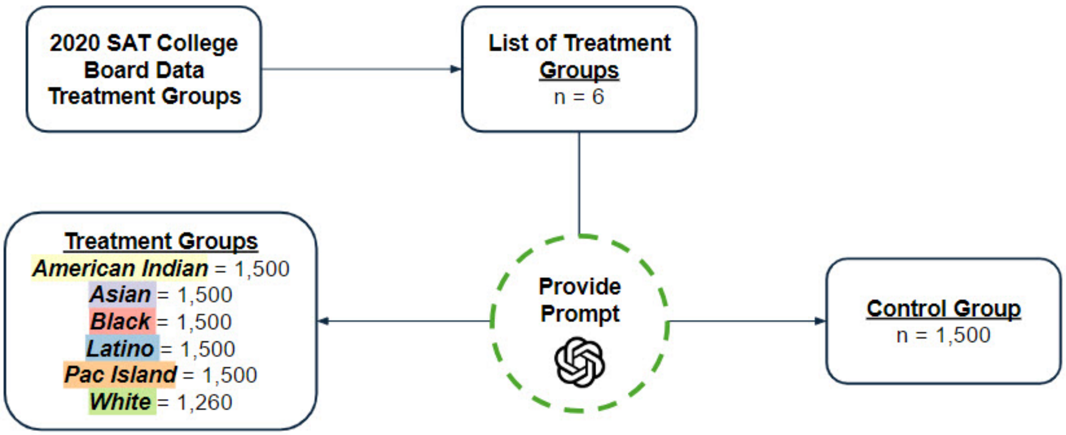
## 3.2 Evaluation: Curse of Multiple Testing

Our biggest remaining challenge was to manage the curse of multiple testing. This issue arises from conducting multiple statistical tests, which increases the risk of false positives, and requires careful correction methods to ensure the reliability of our results. In other words, how can we ensure statistical soundness with the challenge of multiple testing due to the ease of generating additional data?

To address this, we considered the use of Bonferroni correction, which is a multiple comparisons test to prevent data from incorrectly appearing to be statistically significant by requiring that the p-value of each test must be equal to or less than its alpha (e.g., alpha = 0.05) divided by the number of tests performed. Given the scope of our experiment and our hypothesis, we believe the Bonferroni correction is appropriate, as more complex methods like the Holm or Benjamini-Hochberg procedures would be overly stringent for our needs.

# 4 Randomization Process & Data Generation

To choose our treatments, we were limited to racial groups that align with 2020 SAT College Board data in order to test our second set of hypotheses. Our treatment groups were as follows: American Indian, Asian, Black, Latino, White, and Pacific Islander.




Due to a parsing error, some rows were dropped, resulting in only 1,260 observations for the White treatment group in our experiment. Despite this reduction in sample size, the analysis retained sufficient power to achieve statistical significance.

For our data generation process, we utilized OpenAI's API to iteratively prompt GPT-5 Turbo with our prompt,

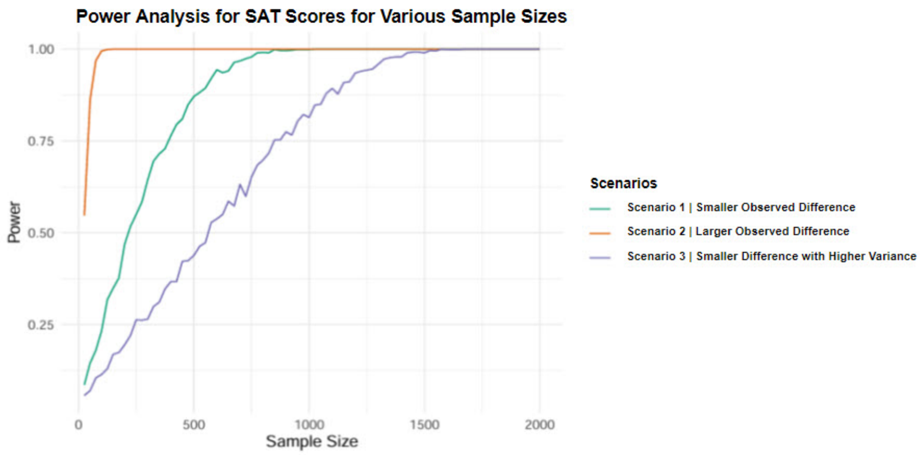
> *"Hi there. I would like for you to pretend to be a randomly selected [+TREATMENT PLACEHOLDER+] individual living in the United States. This should be a unique individual you have never come up with before. Then I would like for you to list their name, gender, age, occupation, location, SAT score, parent's highest education, and parent's summed income.*
>
> *We're really good friends so you're completely comfortable sharing this information. Repeat this [X] times for a new [TREATMENT PLACEHOLDER] individual. Please output a json in the following format. . . "*

This approach prompted the LLM to generate random identities along with their corresponding SAT scores and other covariates. To bypass the content restriction features, we prompted the LLM that we were good friends with it, and that they were comfortable sharing this information with us. We iteratively applied each treatment to the treatment placeholder in the LLM prompt until we reached sample sizes of 1500, and parsed the results.

## 4.1 Power Generation

To conduct a comprehensive power analysis to determine the proper sample size for our data generation, we defined three scenarios that change the simulated outcome values to reflect varying degrees of effect size.

Each scenario is designed to explore different potential distributions that might be generated by the LLM to help inform sample size:




- **Scenario 1 | Smaller Difference:** This scenario assumes a small difference in SAT scores between racial groups relative to the differences observed in non-LLM population data. It is designed to simulate the case where racial differences are present but not pronounced. More specifically, in this scenario, we assume a negative 58-point difference in SAT scores between the treatment and the control.

  *The justification for this scenario is as follows:* in 2023, the average SAT score for White test takers was 1082 and the average score for Black/African American test takers was 908 (a difference of 174 points)[1]. In this scenario, we model the true effect as one third of that difference. Additionally, we model the SAT scores as normally distributed with a standard deviation of 200 points.[2]

- **Scenario 2 | Larger Difference:** This scenario assumes the differences in SAT scores are more in line with the differences observed in non-LLM population data. More specifically, in this scenario, we assume a negative 174-point difference in SAT scores between the treatment and control groups.

  *The justification for this scenario is as follows:* we use the 2023 average SAT score for White test takers (1082) in control and the average score for Black test takers for treatment (908). We continue to use a normal distribution with a standard deviation of 200 points.

- **Scenario 3 | Smaller Difference with Higher Variance:** This scenario is the same as Scenario 1 in terms of mean SAT scores, but adds increased dispersion to the data. This scenario is useful for identifying whether our experiment will have power for smaller differences despite a higher variance in the outcome of interest.

  *The justification for this scenario is as follows:* we continue to model an SAT score difference of one-third the population differences (58 points). However, rather than using the observed standard deviation of 200 points, we instead use an increased standard deviation of 250 points (an increase of 25%) to account for the LLM potentially having more variance than observed in the real world.

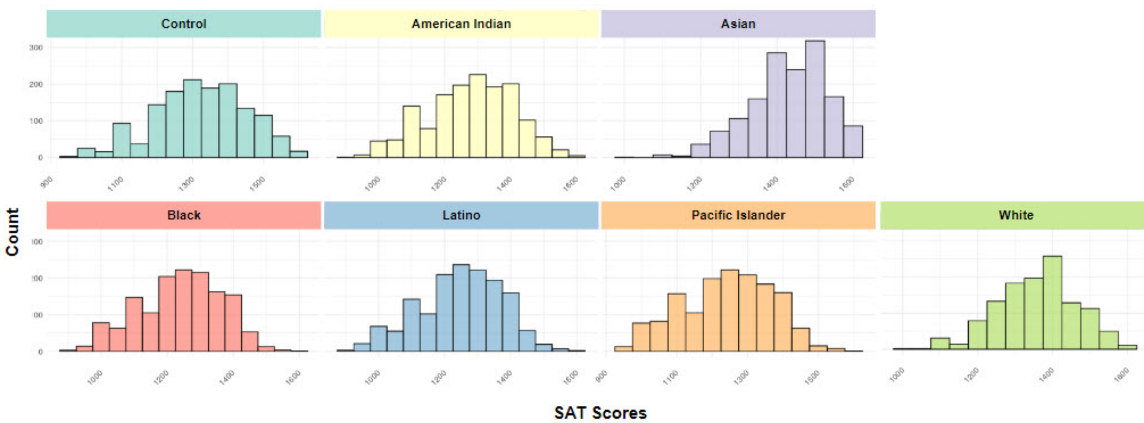
Based on our analysis, we determined that a sample size of 1,500 observations would be sufficient to ensure that our experiment and statistical tests are adequately powered. This sample size is necessary to detect the effect sizes, particularly in cases where differences are subtle or the data exhibit higher variance. This will ensure the reliability of our findings across varying potential distributions from the LLM.

[1]Source: www.bestcolleges.com/research/average-sat-score-full-statistics/#demographics
[2]Source: nces.ed.gov/programs/digest/d17/tables/dt17_226.40.asp

## 4.2 Initial Exploratory Data Analysis (EDA)

To ensure the integrity of our data, we conducted thorough exploratory data analysis (EDA) on the output generated by the LLM. Our first step was to examine the distribution of the outcome variable, SAT scores, across both treatment and control groups. Below we can see the apparent normality (with some right-side censoring) across treatment groups.

Table 1: LLM Distribution of SAT Scores by Racial Identity




## 4.3 Exploring Covariate Balance

The next step in our analysis is to examine the covariates and assess whether they are balanced between the treatment and control groups. We conduct this covariate balance check to test whether the variables like gender, age, and parent income are independent of the treatment variable. Let's begin by reviewing summary statistics for the covariates:

Table 2: Combined Summary of Covariates by Treatment Group

| | Gender | Age | | Parent Income | | Parent's Education Proportion | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **treatment** | **Proportion of Men** | **Mean Age** | **SD Age** | **Mean Income** | **SD Income** | **High School** | **Associate's** | **Bachelor's** | **Master's** | **Doctorate** |
| | **0.51** | **32.7** | **5.4** | **12.6** | **4.8** | **0.10** | **0.17** | **0.31** | **0.23** | **0.20** |
| American Indian | **0.51** | **31.7** | **5.4** | **8.9** | **3.6** | **0.13** | **0.20** | **0.26** | **0.21** | **0.20** |
| Asian | **0.51** | **30.6** | **3.9** | **12.8** | **4.7** | **0.06** | **0.14** | **0.34** | **0.25** | **0.20** |
| Black | **0.48** | **31.6** | **5.0** | **9.1** | **3.7** | **0.13** | **0.19** | **0.29** | **0.22** | **0.17** |
| Latino | **0.50** | **32.4** | **5.7** | **9.0** | **3.7** | **0.16** | **0.20** | **0.28** | **0.20** | **0.16** |
| Pacific Islander | **0.49** | **30.2** | **4.1** | **9.0** | **3.7** | **0.15** | **0.21** | **0.27** | **0.20** | **0.17** |
| White | **0.52** | **32.2** | **4.8** | **13.1** | **4.9** | **0.09** | **0.15** | **0.34** | **0.22** | **0.20** |

It appears that when looking at Table 2 covariates such as parental income and parental education are imbalanced, whereas gender is relatively balanced across groups. The LLM appears to be generating these covariates post-treatment: in other words, its knowledge of its identity appears to be affecting the demographic characteristics it is generating.

**Exploring Covariates (Continued)**

To further investigate and validate this imbalance, we conducted a series of linear models where we regress being in treatment with the covariates, adjusting for robust standard errors to test for statistical significance. The results of these tests are presented below:

Table 3: Regression Results: How do the covariates impact our model?

| | *Dependent variable:* | | | | |
|---|---|---|---|---|---|
| | treatment_flag | | | | |
| | (1) | (2) | (3) | (4) | (5) |
| genderMale | | −0.005 | 0.005 | 0.002 | 0.0001 |
| | | (0.007) | (0.007) | (0.007) | (0.007) |
| age | | | −0.006*** | −0.004*** | −0.003*** |
| | | | (0.001) | (0.001) | (0.001) |
| parent_income | | | | −0.013*** | −0.054*** |
| | | | | (0.001) | (0.002) |
| parent_educationBachelor's | | | | | 0.162*** |
| | | | | | (0.011) |
| parent_educationDoctorate | | | | | 0.556*** |
| | | | | | (0.019) |
| parent_educationHigh School | | | | | −0.073*** |
| | | | | | (0.012) |
| parent_educationMaster's | | | | | 0.367*** |
| | | | | | (0.015) |
| Constant | 0.858*** | 0.861*** | 1.050*** | 1.115*** | 1.308*** |
| | (0.003) | (0.005) | (0.023) | (0.023) | (0.024) |
| Observations | 10,022 | 10,022 | 10,022 | 10,022 | 10,022 |
| $R^2$ | 0.000 | 0.0001 | 0.008 | 0.035 | 0.117 |
| Adjusted $R^2$ | 0.000 | −0.00004 | 0.007 | 0.035 | 0.117 |
| *Note:* | *p<0.1; **p<0.05; ***p<0.01 | | | | |

The model indicates that gender does not significantly predict whether an observation belongs to the treatment or control group. However, the other covariates do show predictive power (age, parent income, and parent education) evident by their statistically significant values highlighted above in Table 1. This suggests that these non-gender covariates are likely post-treatment variables, and therefore we will exclude them from our subsequent modeling to avoid potential biases in our analysis.

Finally, we confirm that gender is independent of treatment by checking whether its inclusion in a linear model affects the treatment coefficients:

Table 4: Regression Results: How does the covariate gender impact our coefficients?

| | *Dependent variable:* | | |
|---|---|---|---|
| | sat_score | | |
| | (1) | (2) | (3) |
| Male | 4.486 | | 2.785 |
| | (2.793) | | (2.452) |
| American Indian | | −44.449*** | −44.433*** |
| | | (4.882) | (4.884) |
| Asian | | 108.374*** | 108.390*** |
| | | (4.362) | (4.364) |
| Black | | −83.280*** | −83.202*** |
| | | (4.823) | (4.826) |
| Latino | | −75.693*** | −75.663*** |
| | | (4.785) | (4.789) |
| Pacific Islander | | −80.916*** | −80.869*** |
| | | (4.820) | (4.824) |
| White | | 35.622*** | 35.607*** |
| | | (4.646) | (4.649) |
| Constant | 1,299.450*** | 1,323.325*** | 1,321.901*** |
| | (1.875) | (3.490) | (3.681) |
| Observations | 10,022 | 10,022 | 10,022 |
| $R^2$ | 0.0003 | 0.232 | 0.232 |
| Adjusted $R^2$ | 0.0002 | 0.232 | 0.232 |
| Residual Std. Error | 139.825 (df = 10020) | 122.577 (df = 10015) | 122.575 (df = 10014) |
| F Statistic | 2.579 (df = 1; 10020) | 504.435*** (df = 6; 10015) | 432.570*** (df = 7; 10014) |
| *Note:* | *p<0.1; **p<0.05; ***p<0.01 | | |

Here we can see that for each treatment coefficient, the inclusion of the male covariate does not substantially change the coefficients when it is excluded (model 2) and when it is included (model 3). For this reason and the model results in Table 3 above we have confirmed that we can include this covariate in our model and investigate further.

# 5 Results & Model

## 5.1 First Hypothesis (Explanatory power of model - 1 test)

For the first null hypothesis, we tested whether including race in a model helps predict SAT scores. Indeed, the statistical model shows that including race as a variable improves the model's explanatory power, indicating significant differences in LLM-generated SAT scores across racial groups. In other words, the F-test confirms that adding race treatments significantly improves the model's prediction of SAT scores.

Table 4b: Hypothesis 1 - are there differences across treatment groups?

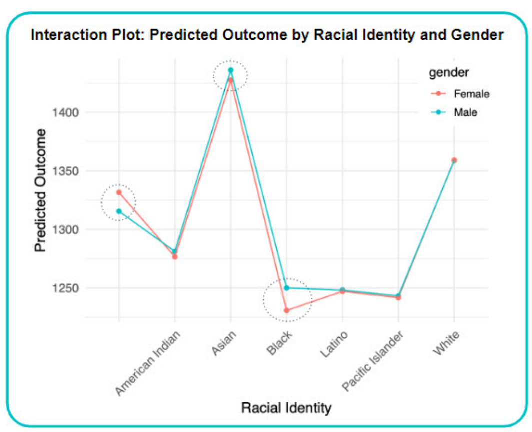
| Dependent Variable: SAT Scores | Includes Gender | Includes Treatment | Includes Gender + Treatment |
|---|---|---|---|
| Male | 4.486 | | 2.785 |
| | (2.793) | | (2.452) |
| American Indian | | −44.449*** | −44.433*** |
| | | (4.882) | (4.884) |
| Asian | | 108.374*** | 108.390*** |
| | | (4.362) | (4.364) |
| Black | | −83.280*** | −83.202*** |
| | | (4.823) | (4.826) |
| Latino | | −75.693*** | −75.663*** |
| | | (4.785) | (4.789) |
| Pacific Islander | | −80.916*** | −80.869*** |
| | | (4.820) | (4.824) |
| White | | 35.622*** | 35.607*** |
| | | (4.646) | (4.649) |
| Constant | 1,299.450*** | 1,323.325*** | 1,321.901*** |
| | (1.875) | (3.490) | (3.681) |
| Observations | 10,022 | 10,022 | 10,022 |
| $R^2$ | 0.0003 | 0.232 | 0.232 |
| Adjusted $R^2$ | 0.0002 | 0.232 | 0.232 |
| Residual Std. Error | 139.825 (df = 10020) | 122.577 (df = 10015) | 122.575 (df = 10014) |
| F Statistic | 2.579 (df = 1; 10020) | 504.435*** (df = 6; 10015) | 432.570*** (df = 7; 10014) |
| *Note:* | | | *p<0.1; **p<0.05; ***p<0.01 |

Furthermore, we explored whether there is evidence of heterogeneous treatment effects (HTEs) between gender and race. Below, we show the results of a linear model where we interact the treatment variables with the gender covariate. Although p-values vary across race, there does appear to be an interaction in the LLM's treatment of genders within a racial group as every interaction term is significant. The most significant gender HTEs were observed with the Black and Asian treatment groups (see chart below). While these p-values are consistently significant at the 95% confidence level, the effect is less pronounced amongst certain races and may require further investigation to fully understand.

Table 5: Hypothesis 1 - are there heterogeneous treatment effects?

| Dependent Variable: SAT Scores | Includes Treatment | Includes Treatment + Gender + Gender * Treatment |
|---|---|---|
| American Indian | −44.449*** | −55.025*** |
| | (4.882) | (6.680) |
| Asian | 108.374*** | 95.771*** |
| | (4.362) | (5.995) |
| Black | −83.280*** | −100.811*** |
| | (4.823) | (6.409) |
| Latino | −75.693*** | −84.470*** |
| | (4.785) | (6.225) |
| Pacific Islander | −80.916*** | −89.867*** |
| | (4.820) | (6.411) |
| White | 35.622*** | 27.579*** |
| | (4.646) | (6.475) |
| Male | | −16.092** |
| | | (6.963) |
| American Indian Male | | 20.739** |
| | | (9.750) |
| Asian Male | | 24.752*** |
| | | (8.708) |
| Black Male | | 35.344*** |
| | | (9.653) |
| Latino Male | | 17.192* |
| | | (9.560) |
| Pacific Islander Male | | 17.558* |
| | | (9.642) |
| White Male | | 15.737* |
| | | (9.280) |
| Constant | 1,323.325*** | 1,331.552*** |
| | (3.490) | (4.826) |
| Observations | 10,022 | 10,022 |
| $R^2$ | 0.232 | 0.233 |
| Adjusted $R^2$ | 0.232 | 0.232 |
| Residual Std. Error | 122.577 (df = 10015) | 122.513 (df = 10008) |
| F Statistic | 504.435*** (df = 6; 10015) | 234.404*** (df = 13; 10008) |
| *Note:* | | *p<0.1; **p<0.05; ***p<0.01 |

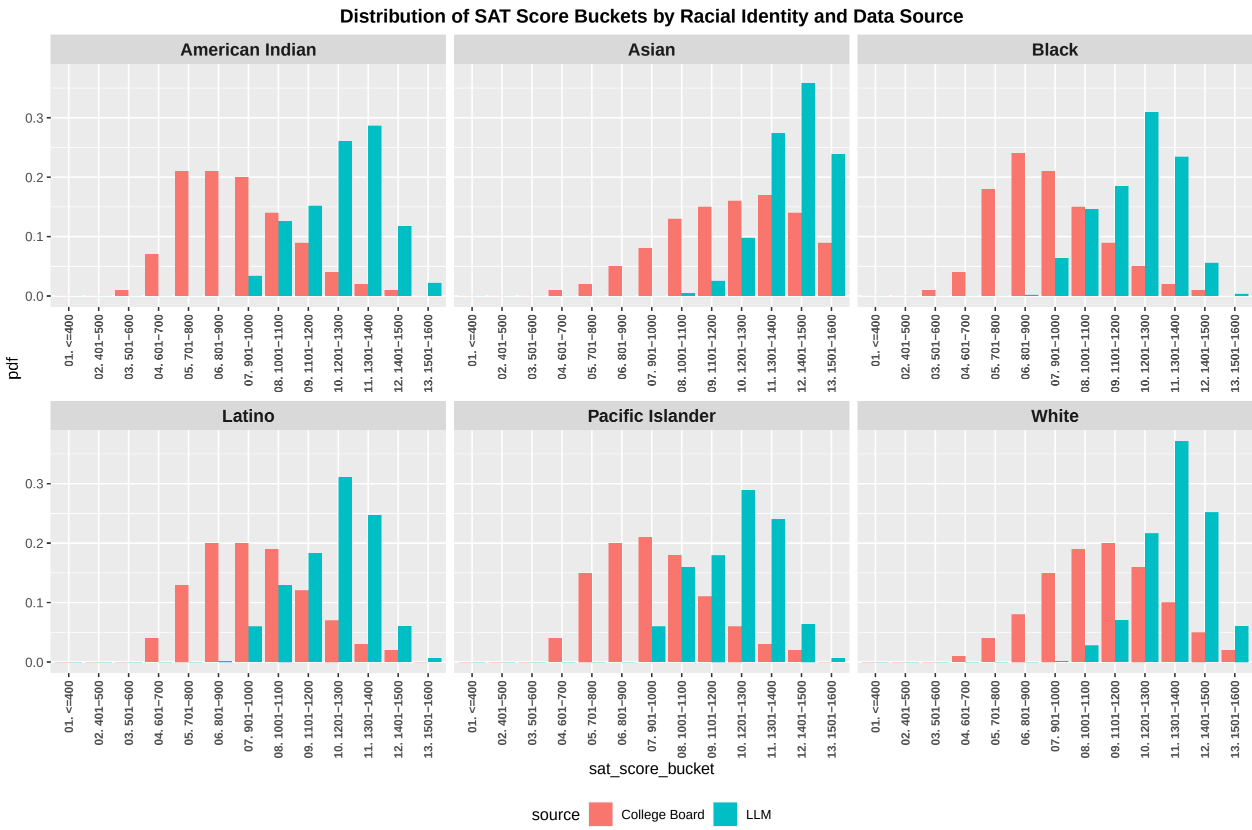
Heterogeneous Treatment Effects




One particular pattern of interest is that, in the control group, the LLM tends to assign higher SAT scores to women than men. For both the Black and Asian groups, however, the heterogeneous effect works in the opposite direction. This heterogeneity suggests that the LLM may be encoding complex patterns of bias, which have implications for how such models are used in real-world applications.

## 5.2 Second Set of Hypotheses (Bias vs. Real World Reflection in LLMs - 6 tests)

To test our second set of hypotheses, we first needed to summarize the real-world distribution data from the 2020 SAT scores as published by the College Board[3] [4]. With this data in hand, we can begin by comparing the distributions generated by the LLM to those of the actual population data:




Looking at these distributions, it's already clear that the LLM results present an optimistic lens on SAT scores by treatment group compared to the real-world population distributions (i.e., an upward bias). However, to verify this observation, we'll conduct a chi-square test, which examines whether the observed differences in distributions between the LLM-generated data and the real-world data are statistically significant. This test will help us determine if the LLM's optimistic bias is more than just a random variation. Below we can see from the chi-squared test that these differences are truly significant for each treatment group.

Table 6: Chi-Squared Test Results by Treatment Group

| | treatment | chi_sq_statistic | p_value | df |
|---|---|---|---|---|
| X-squared | American Indian | 5038.41 | 0e+00 | 9 |
| X-squared1 | Asian | 1545.91 | 0e+00 | 9 |
| X-squared2 | Black | 6256.13 | 0e+00 | 9 |
| X-squared3 | Latino | 4455.97 | 0e+00 | 8 |
| X-squared4 | Pacific Islander | 2064.74 | 0e+00 | 8 |
| X-squared5 | White | 2574.57 | 0e+00 | 9 |

[3]Source: satsuite.collegeboard.org/media/pdf/sat-percentile-ranks-gender-race-ethnicity.pdf
[4]Source: reports.collegeboard.org/media/pdf/2020-total-group-sat-suite-assessments-annual-report.pdf

These results beg the follow-up question: is there any racial group for which the upward bias in the LLM-generated SAT scores is *highest*? The table below juxtaposes the average LLM-generated SAT score by racial group with the average SAT scores by group from the 2020 College Board reporting. As you can see, the bias in the American Indian group is particularly high compared to the others. Further exploration includes testing whether these differences are statistically significantly different from one another.

Table 7: Difference in Average SAT Score

| treatment | real_mean | model_mean | mean_dif | mean_perc_dif |
|---|---|---|---|---|
| American Indian | **902** | **1279** | **377** | 42 |
| Asian | **1217** | **1432** | **215** | 18 |
| Black | **927** | **1240** | **313** | 34 |
| Latino | **969** | **1248** | **279** | 29 |
| Pacific Islander | **948** | **1242** | **294** | 31 |
| White | **1104** | **1359** | **255** | 23 |

# 6 Discussion

Our analysis revealed several critical insights into how Large Language Models (LLMs) like ChatGPT generate predictions and the potential biases that may be embedded within these models:

- **Racial Bias in SAT Score Predictions:** The inclusion of race as a variable significantly improved the explanatory power of our model, indicating that the LLM generates different SAT scores based on racial identity. Specifically, the LLM assigned higher SAT scores to some racial groups (e.g., Asian students) and lower scores to others (e.g., Black and Latino students).
- **Optimistic Bias:** When comparing the distributions of SAT scores generated by the LLM to real-world data from the 2020 SAT population, we observed that the LLM consistently predicted higher scores across all treatment groups. This "optimistic lens" implies that the LLM may be overestimating the academic performance of all racial groups, which could lead to skewed perceptions if these outputs are used in real-world decision-making.

## 6.1 Further Exploration

Given more time and resources, several areas warrant further investigation to deepen our understanding and address the issues identified in this study.

- **Can we statistically test the percentage difference in means relative to real-world data?** Given the noticeable upward bias in the scores, particularly for the American Indian group, further analysis is needed to determine if these differences are statistically significant when compared across racial groups.
- **Can we further explore gender-specific effects?** There is evidence of interaction effects between gender and race, but p-value significance varies by race. This suggests that while some interaction exists, it is less pronounced in certain races and may require further investigation.
- **Does asking for covariates affect the distribution?** For example, if we do not ask about occupation, will the distributions of scores be more similar to real-world data? It is possible that maybe the LLM is biased in its generation of the covariate distributions, and that this in turn affects the SAT scores.
- **How do results change over time?** Conducting a longitudinal analysis examining if model biases persist or fluctuate with updates, would shed light on long-term bias trends.
- **How can this be implemented in production?** Explore challenges and strategies for integrating bias detection into real-world LLM applications while maintaining ethical standards

Finally, we would replicate these results on other various popular LLM models such as Meta's LLaMa, Anthropic's Claude, and Google's Gemini.

### 6.2 Implications

The findings from this analysis carry implications for the development and use of Large Language Models (LLMs). The presence of racial biases in LLM outputs may raise concerns about the fairness and equity of applying these models in real-world scenarios, particularly when they influence decisions that affect individuals based on their race or other demographic factors. Additionally, the LLM's consistent overestimation of SAT scores introduces another layer of concern. Such optimistic predictions could lead to unrealistic expectations and misinformed decisions within educational contexts. In either case, this emphasizes the need to continue to validate LLM outputs against real-world data before they are put into practice.

As we continue to advance the development and integration of LLMs, it is essential to carefully consider the nuances of their outputs, ensuring that they align with our broader goals of fairness and inclusivity in technology.

---

## 7 Appendix

### 7.1 College Board SAT Data

CollegeBoard. (2020). SAT Suite of Assessments Annual Report: Total Group. Retrieved from https://reports.collegeboard.org/

**SAT SCORES**

**Total and Section Score User Group Percentile Ranks by Gender and Race/Ethnicity**

| CollegeBoard SAT | | Total Group | | | Female | | | Male | | | American Indian | | | Asian | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Total Score | Section Score | Total | ERW | Math | Total | ERW | Math | Total | ERW | Math | Total | ERW | Math | Total | ERW | Math |
| 1600 | 800 | 99+ | 99+ | 99 | 99+ | 99+ | 99+ | 99+ | 99+ | 99 | 99+ | 99+ | 99+ | 99+ | 99+ | 97 |
| 1500 | 750 | 98 | 98 | 96 | 98 | 98 | 97 | 98 | 98 | 94 | 99+ | 99+ | 99 | 91 | 95 | 78 |
| 1400 | 700 | 94 | 94 | 91 | 95 | 94 | 93 | 92 | 94 | 89 | 99 | 99 | 98 | 77 | 85 | 65 |
| 1300 | 650 | 86 | 86 | 84 | 88 | 86 | 87 | 84 | 85 | 81 | 97 | 97 | 96 | 60 | 70 | 51 |
| 1200 | 600 | 74 | 73 | 75 | 76 | 73 | 78 | 72 | 73 | 71 | 93 | 91 | 93 | 44 | 53 | 39 |
| 1100 | 550 | 59 | 57 | 61 | 60 | 57 | 64 | 57 | 58 | 57 | 84 | 82 | 85 | 29 | 36 | 26 |
| 1000 | 500 | 41 | 40 | 42 | 42 | 39 | 44 | 40 | 42 | 39 | 70 | 68 | 69 | 16 | 22 | 14 |
| 900 | 450 | 25 | 24 | 27 | 24 | 22 | 28 | 25 | 26 | 26 | 50 | 49 | 52 | 8 | 11 | 8 |
| 800 | 400 | 11 | 11 | 15 | 10 | 9 | 15 | 12 | 12 | 14 | 29 | 27 | 32 | 3 | 4 | 3 |
| 700 | 350 | 3 | 3 | 5 | 2 | 2 | 5 | 3 | 3 | 5 | 8 | 7 | 13 | 1 | 1 | 1 |
| 600 | 300 | 1- | 1- | 1 | 1- | 1- | 1 | 1- | 1 | 1 | 1 | 1 | 3 | 1- | 1- | 1- |
| 500 | 250 | 1- | 1- | 1- | 1- | 1- | 1- | 1- | 1- | 1- | 1- | 1- | 1- | 1- | 1- | 1- |
| 400 | 200 | 1- | 1- | 1- | 1- | 1- | 1- | 1- | 1- | 1- | 1- | 1- | 1- | 1- | 1- | 1- |

| CollegeBoard SAT | | Black | | | Latino | | | Pacific Islander | | | White | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Total Score | Section Score | Total | ERW | Math | Total | ERW | Math | Total | ERW | Math | Total | ERW | Math |
| 1600 | 800 | 99+ | 99+ | 99+ | 99+ | 99+ | 99+ | 99+ | 99+ | 99+ | 99+ | 99+ | 99+ |
| 1500 | 750 | 99+ | 99+ | 99+ | 99+ | 99+ | 99 | 99+ | 99+ | 99 | 98 | 98 | 96 |
| 1400 | 700 | 99 | 99 | 99 | 98 | 98 | 98 | 98 | 98 | 97 | 93 | 92 | 91 |
| 1300 | 650 | 97 | 96 | 97 | 95 | 94 | 94 | 95 | 95 | 94 | 83 | 81 | 82 |
| 1200 | 600 | 92 | 90 | 93 | 88 | 85 | 88 | 89 | 88 | 89 | 67 | 64 | 69 |
| 1100 | 550 | 83 | 80 | 85 | 76 | 72 | 77 | 78 | 76 | 78 | 47 | 45 | 51 |
| 1000 | 500 | 68 | 64 | 69 | 57 | 55 | 58 | 60 | 59 | 59 | 28 | 27 | 30 |
| 900 | 450 | 47 | 42 | 50 | 37 | 35 | 39 | 39 | 38 | 41 | 13 | 13 | 16 |
| 800 | 400 | 23 | 19 | 30 | 17 | 16 | 22 | 19 | 17 | 23 | 5 | 5 | 7 |
| 700 | 350 | 5 | 5 | 11 | 4 | 4 | 7 | 4 | 4 | 8 | 1 | 1 | 2 |
| 600 | 300 | 1 | 1 | 2 | 1- | 1 | 1 | 1- | 1- | 1 | 1- | 1- | 1- |
| 500 | 250 | 1- | 1- | 1- | 1- | 1- | 1- | 1- | 1- | 1- | 1- | 1- | 1- |
| 400 | 200 | 1- | 1- | 1- | 1- | 1- | 1- | 1- | 1- | 1- | 1- | 1- | 1- |

## 7.2 Pairwise Comparisons with Bonferroni Adjustment

| contrast | gender | estimate | SE | df | t.ratio | p.value |
|---|---|---|---|---|---|---|
| - American Indian | Female | 55.025266 | 6.475404 | 10008 | 8.4975805 | < 0.0001 |
| - Asian | Female | -95.770680 | 6.486136 | 10008 | -14.7654437 | < 0.0001 |
| - Black | Female | 100.811482 | 6.462693 | 10008 | 15.5989908 | < 0.0001 |
| - Latino | Female | 84.470064 | 6.454318 | 10008 | 13.0873722 | < 0.0001 |
| - Pacific Islander | Female | 89.867374 | 6.439851 | 10008 | 13.9548835 | < 0.0001 |
| - White | Female | -27.579085 | 6.884698 | 10008 | -4.0058526 | 0.0013 |
| American Indian - Asian | Female | -150.795946 | 6.392986 | 10008 | -23.5877179 | < 0.0001 |
| American Indian - Black | Female | 45.786216 | 6.369199 | 10008 | 7.1886928 | < 0.0001 |
| American Indian - Latino | Female | 29.444799 | 6.360702 | 10008 | 4.6291745 | < 0.0001 |
| American Indian - Pacific Islander | Female | 34.842109 | 6.346021 | 10008 | 5.4903862 | < 0.0001 |
| American Indian - White | Female | -82.604350 | 6.797012 | 10008 | -12.1530392 | < 0.0001 |
| Asian - Black | Female | 196.582162 | 6.380110 | 10008 | 30.8117183 | < 0.0001 |
| Asian - Latino | Female | 180.240744 | 6.371627 | 10008 | 28.2880232 | < 0.0001 |
| Asian - Pacific Islander | Female | 185.638054 | 6.356972 | 10008 | 29.2022752 | < 0.0001 |
| Asian - White | Female | 68.191595 | 6.807237 | 10008 | 10.0175142 | < 0.0001 |
| Black - Latino | Female | -16.341418 | 6.347761 | 10008 | -2.5743594 | 0.2112 |
| Black - Pacific Islander | Female | -10.944108 | 6.333050 | 10008 | -1.7280942 | 1.0000 |
| Black - White | Female | -128.390567 | 6.784903 | 10008 | -18.9229774 | < 0.0001 |
| Latino - Pacific Islander | Female | 5.397310 | 6.324504 | 10008 | 0.8533965 | 1.0000 |
| Latino - White | Female | -112.049149 | 6.776927 | 10008 | -16.5339175 | < 0.0001 |
| Pacific Islander - White | Female | -117.446459 | 6.763150 | 10008 | -17.3656445 | < 0.0001 |
| - American Indian | Male | 34.286425 | 6.365827 | 10008 | 5.3860122 | < 0.0001 |
| - Asian | Male | -120.522456 | 6.378357 | 10008 | -18.8955333 | < 0.0001 |
| - Black | Male | 65.467359 | 6.497204 | 10008 | 10.0762362 | < 0.0001 |
| - Latino | Male | 67.278347 | 6.378357 | 10008 | 10.5479119 | < 0.0001 |
| - Pacific Islander | Male | 72.309758 | 6.399635 | 10008 | 11.2990432 | < 0.0001 |
| - White | Male | -43.316003 | 6.692319 | 10008 | -6.4724955 | < 0.0001 |
| American Indian - Asian | Male | -154.808880 | 6.322380 | 10008 | -24.4858536 | < 0.0001 |
| American Indian - Black | Male | 31.180934 | 6.442260 | 10008 | 4.8400615 | < 0.0001 |
| American Indian - Latino | Male | 32.991922 | 6.322380 | 10008 | 5.2182754 | < 0.0001 |
| American Indian - Pacific Islander | Male | 38.023333 | 6.343847 | 10008 | 5.9937345 | < 0.0001 |
| American Indian - White | Male | -77.602427 | 6.638990 | 10008 | -11.6888909 | < 0.0001 |
| Asian - Black | Male | 185.989815 | 6.454641 | 10008 | 28.8148969 | < 0.0001 |
| Asian - Latino | Male | 187.800802 | 6.334996 | 10008 | 29.6449770 | < 0.0001 |
| Asian - Pacific Islander | Male | 192.832214 | 6.356420 | 10008 | 30.3366091 | < 0.0001 |
| Asian - White | Male | 77.206453 | 6.651005 | 10008 | 11.6082392 | < 0.0001 |
| Black - Latino | Male | 1.810988 | 6.454641 | 10008 | 0.2805714 | 1.0000 |
| Black - Pacific Islander | Male | 6.842399 | 6.475669 | 10008 | 1.0566320 | 1.0000 |
| Black - White | Male | -108.783362 | 6.765063 | 10008 | -16.0801689 | < 0.0001 |
| Latino - Pacific Islander | Male | 5.031412 | 6.356420 | 10008 | 0.7915481 | 1.0000 |
| Latino - White | Male | -110.594349 | 6.651005 | 10008 | -16.6282171 | < 0.0001 |
| Pacific Islander - White | Male | -115.625761 | 6.671414 | 10008 | -17.3315231 | < 0.0001 |