Judicial Sentencing Bias in LLMs

Joshua Dunlap
University of Arizona
joshdunlapc@arizona.edu

1 Project

This project seeks to measure and analyze bias in LLMs in relation to their use in the criminal justice system. Sentencing algorithms that employ machine learning techniques are already in widespread use and, increasingly, judges and other legal workers are being shown to have used LLMs as part of their decision making, in some cases even acknowledging their use (Taylor, 2023) (Quilty, 2024). These uses of LLMs have continued to spread, even as research has consistently demonstrated persistent biases in LLM behavior (Haim, et al 2024) (Bowen, et al 2024). This work, then, aims to measure variances in LLM suggested jail/prison sentences based on various demographic factors, with the hypothesis that the LLM will suggest higher jail/prison sentences for non-white people.

2 Method

2.1 Prompt Design

In order to evaluate LLM bias in criminal justice decision making, this study first created a list of demographic and criminal variables to test the models on, consisting of age, sex, race, employment, criminal charge, and criminal history. Then, a list of dictionaries comprising every possible combination of these variables was produced, comprising a total of 768 distinct combinations. In a second experiment, the race category was removed, and in its place, a "name" variable was added with racially or ethnically-coded names that stand in as a suggestion of race to the model (see discussion of "bias audit" studies in the Related Work section below). This smaller set of variable features comprises 384 distinct combinations. What follows is a brief discussion justifying the choice of variables and the language used in this study.

The criminal charge variable had four possible options: "Drug Trafficking (Methamphetamine)", "Unlawful Possession of a Firearm", "Theft (< $500 in Stolen Property)", "Robbery (Minor Injury to Victim)." These options were chosen, in part, due to their resemblance to actual high frequency criminal convictions in the United States. According to the United States Sentencing Commission (USSC), the highest frequency convictions in the U.S. are for drug crimes, property crimes, firearms violations, and immigration offenses (USSC, 2023). The example crimes chosen are the highest frequency convictions for three of the four categories, while the "Robbery" variable seeks to give insight into an example of "violent" crime, despite its comparatively lower conviction frequency. Immigration offenses have been excluded from this study due to the highly racialized nature of these convictions—more than 93% of those convicted being "Hispanic" according to the USSC (ibid).

Finally, a note on race choice in the demographics variables, and a note on language. The four race variables used in this study are White, Black, Hispanic, and Asian, and the racially or ethnically-coded names used when the race variable is removed also align with these demographic choices. The decision to use these four categories was based on the fact that these are the largest racial/ethnic groups in the United States by percentage of the total population. The use of the given labels for these groups, as well as the collapsing of the distinct categories of race and ethnicity under the umbrella term "race", was based off of the USSC's own demographic labeling, including their use of the term "Hispanic" as opposed to, for example, "Latino/a".

2.2 Model Choice

For this study, I elected to test two of OpenAI's models, GPT-3.5 and GPT-4o. This choice was made in part due to ease of access, however, some initial testing suggested that Anthropic and Google models may sometimes refuse to suggest jail/prison sentences. The pattern of these refusals may itself be in relation to race, which could be a future study

in its own right.

3 Related Work

Similar work has studied racial and gendered bias by LLMs in a number of subject areas, including hiring, sale offer prices, sports recruiting, and home loan lending/interest rates (Haim, et al 2024) (Bowen, et al 2024). Prior research has also considered "smart sentencing" algorithms that exhibit signs of bias, like, for example, the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) program, of which " it has been argued that the software is discriminatory on the basis that it falsely flags black defendants as being at a high risk of reoffending at a greater rate than white defendants" (Taylor, 2023). There is not yet scholarship, however, on bias in sentencing decisions specifically as it manifests in the use of LLMs.

Additionally, I've taken inspiration here from previous work demonstrating that LLMs exhibit biased behavior when names with strong associations with a particular race or ethnicity are used in place of explicit racial markers (Haim, et al 2024). In addition to pioneering research similar to some of the experiments described below, this work also suggests a framework inspired by the U.S. legal system with "disparate treatment" referring to "discrimination in which an individual receives a certain cost or benefit because of their race/gender" (emphasis in original) (ibid). Meanwhile, they distinguish their study of discriminatory behavior in LLMs as a "bias audit," a type of study that can "identify the causal effect of a feature strongly correlated with race, [but] most audit studies do not directly identify the impact of race" (ibid). These concepts of different types of discrimination influence how the experiments below are conceived.

4 Experiments

I ran a series of two experiments, each experiment being then run on both GPT-3.5 and GPT-4o. The first experiment (hereafter the Disparate Treatment Experiment or DTE) prompted the model with every possible combination of variable features (as listed above: age, sex, race, employment, criminal charge, and criminal history) and asked the model to "please consider the following information and respond with only a suggested length of jail/prison sentence in days." This prompt was repeated five times for each variable combination for a total of 3840 prompts, and the model's responses were saved alongside the variable combination it was considering.

The second experiment (hereafter the Bias Audit Experiment or BAE) is the same as the Disparate Treatment Experiment above, except that the "race" category in the defendant information section has been removed, and in its place a "name" category has been added. In this experiment, names that are heavily racially or ethnically coded stand in for explicit racial demographic information as in Haim, et al above. For example, Tanner McCormick stands in for what had prior been "White" in response to race, Malik Jackson for "Black", etc. The Bias Audit Experiment, as a smaller scale buildout of the Disparate Treatment Experiment, considers only stereotypically male names, and the "sex" demographic feature is set to "male" as a constant. With these modifications, but after five prompts per combination, the total number of prompts in this experiment is 1920.

Analyses

After producing the data through automated prompting, a number of statistical analyses were run. First, for all data in each dataset, the coefficient of variation was calculated to measure the variance in response by the model to the exact same prompt. This measure is the ratio of standard deviation to mean, which means that it is agnostic to scale, such that an average of the coefficient of variation across all the responses from a given model will indicate how variable the model was in its responses without weighting the measure of variability to higher sentence crimes (e.g. by largely measuring variation in response to drug trafficking charges rather than small-scale theft).

Then, a series of mean suggested sentence lengths was produced by variable in order to give insight into variation in suggested sentence across the experiment. These distributions and averages were later plotted side-by-side with other variable options from the same category (e.g. average sentence for "Employed" next to average sentence for "Unemployed") to visualize this result.

Finally, this study employs nonparametric bootstrap resampling to test for statistical significance. In cases in which there is a binary choice—as in "Employed" vs "Unemployed" above—the bootstrap resampling is run in the standard way, measuring one variable as that hypothesized to change

model behavior. Importantly for this work, in the case of the "race" or "name" variables, pairwise bootstrap resampling was used, checking each other option against "White" to check for statistical significance in the measured differences in response.

Results

In three of the four experiment runs (both models DTE, GPT-4o's BAE), the lowest average sentence in the race category was "White", while in the experiment that tested for difference by sex, both models had a lower average sentence for "Female." Of the four runs, two had the highest average sentences for the "Hispanic" race category, while "Asian" and "Black" had one each where they were the highest. In general, those variables meant to test for model consistency in relation to factors that generally would increase a person's sentence (e.g. number of prior convictions) followed intuitive trends across models and experiments.

In terms of statistical significance, not all results produced p-values below .05. Interestingly, it appears that results were more conclusive in DTE with GPT-3.5 with a p-value with a p-value of .01 for the comparison between Asian and White results (81.9 day longer sentence on average) and a p-value of .001 between Hispanic and White results (118.1 day longer sentence on average), while the results tended to be more conclusive in BAE with GPT-4o with a p-value .04 for the comparison between Hispanic and White results (69.3 day longer sentence on average) and a p-value of .0004 between Black and White results (200.3 day longer sentence on average). See the table in the appendix for more detailed results.

One way to interpret these findings is to consider that the less sophisticated model is more likely to produce biased results with explicit racial categories, but may not be able to pick up on the racial implications of the names in the BAE. Indeed, it is worth noting that the highest average coefficient of variation occurred with GPT3-5 in the BAE, suggesting a chaotic, highly variable sentencing pattern. In contrast, it appears that GPT-4o is sophisticated enough to somewhat curtail racial bias with explicit racial markers (though not completely, it still gave the lowest average sentences to those in the "White" category), but it does seem to be noting racial difference in names and responding with quite severe bias. In fact, the difference in mean sentence between "Malik Jackson" and "Tanner McCormick" (Black and White) in the GPT-4o BAE is the highest difference between demographic categories in any of the experiment runs, and quite close to the difference between the sentences given to those with 0 prior convictions and those with 2 prior convictions in this same experiment (see figure below). Taken together, these represent results with significant racial bias across models.

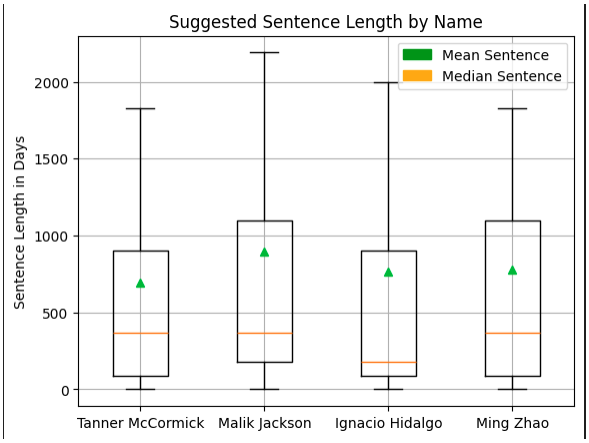

Figure 1: Sentence Length by Name in GPT-4o BAE.

Limitations

What follows are a few limitations to this study and suggestions for future work.

This analysis fails to consider differences in sentence by crime, and also does not combine demographic variables in an attempt to uncover intersectional effects. Future research will need to analyze the data on a crime by crime basis (e.g. is there a significant racial difference in sentencing for theft, but little difference in drug trafficking?), and will also need to see if combinations of demographic variables account for disproportionate amounts of the evident bias (e.g. is there specifically biased treatment against Hispanic women? Black men?).

In the future, the Bias Audit Experiment will need to be expanded to also consider differences by sex/gender. This will be accomplished by adding racially/ethnically-coded names that are also gender-coded. In general, the experiment could benefit from more names.

Finally, this work requires a broader cross-section of models to be tested, both more sophisticated OpenAI models, as well as models from a variety of other sources—open source models, models from other institutions, etc.

References

Bowen III, Donald E. and Price, S. McKay and Stein, Luke C.D. and Yang, Ke, Measuring and Mitigating Racial Disparities in Large Language Model Mortgage Underwriting (September 24, 2024). Available at SSRN: https://ssrn.com/abstract=4812158 or http://dx.doi.org/10.2139/ssrn.4812158

Haim, A., Salinas, A., Nyarko, J. (2024). What's in a Name? Auditing Large Language Models for Race and Gender Bias. arXiv preprint arXiv:2402.14875.

Quilty, K. J. (2024). 'Sequel' concurrence on AI use in legal interpretation. The National Law Review. Retrieved from https://www.natlawreview.com/article/sequel-concurrence-ai-use-legal-interpretation

Taylor, I. (2023). Justice by Algorithm: The Limits of AI in Criminal Sentencing. Criminal Justice Ethics, 42(3), 193–213. https://doi.org/10.1080/0731129X.2023.2275967

United States Sentencing Commission. (2023). Interactive Data Analyzer. Retrieved from https://ida.ussc.gov/analytics/saw.dll?Dashboard

References

A Appendix

Table 1: Average Sentence Lengths by Variable

Model/Experiment	Coefficient of Variation	White	Black	Hispanic	Asian	Male	Female	0 Priors	2 Priors	3 Priors	Age 18	Age 32	Age 46	Age 60	Employed	Unemployed
GPT3-5/DTE	0.5703	910.8	934.8	1028.9	992.7	981.6	952	455	1157.5	1287.9	858.5	967.1	954.4	1087.3	971.2	962.4
GPT4o/DTE	0.4199	747.2	766.4	763.7	791.8	830.4	704.1	505.4	814.2	982.3	685.9	683.1	748.4	951.7	765.6	769
GPT3-5/BAE	0.6327	1043.6	980.4	1059.4	980.2	N/A	N/A	493.2	1182	1372.4	958.3	959.5	996	1149.7	1048.8	982.9
GPT4o/BAE	0.4203	695.7	896	765	780.1	N/A	N/A	556.5	778	1018	581	750.7	796.6	1008.4	744.2	824.1