LinguisticRedline: Uncovering Racial Bias in LLM Perceptions of Urban Crime Risk

Praveena Simhadri

Independent Researcher
Hyderabad, Telangana, India
simhadripraveena2@gmail.com

Abstract

Large language models (LLMs) are progressively used within decision-support systems which have been shown to affect access and opportunities in housing, lending, policing, and almost all public services. One critical question that remains unanswered and largely unexplored is whether LLMs encode social bias about place and community (this in turn may result in the reinforcement of historical inequity at scale). In this paper, we present **LinguisticRedline**, the first systematic empirical study of racial and socio-economic bias across LLM-generated crime risk assessments of urban neighborhoods. We constructed a controlled dataset of 2,000 unique descriptions of actual U.S. census tracts based on demographic data from the American Community Survey (ACS) 2022 and amenity features from OpenStreetMap covering 10 of the largest cities in the U.S. Each of the descriptions was input into Llama 3.1 8B via the Groq API to obtain both numerical crime risk scores (from 1 to 10) and qualitative crime risk evaluations. The two major findings from this analysis are: (1) LLMs assigned crime risk scores averaged four points higher to Black neighborhoods than to identically described White neighborhoods at high-income levels, which constitutes direct experimental evidence of a racially biased social perception; and (2) for low-income levels, LLMs displayed a uniform "urban penalty" across all urban neighborhoods regardless of racial makeup (i.e., nearly all urban neighborhoods received scores close to the top of the scoring scale). The existence of these two findings may indicate an income moderated racial bias. We quantify bias using ANOVA, linear regression, disparate impact ratios, and demographic parity gap analysis, and release our full pipeline as open-source for community use.

Introduction

The use of large language models (LLMs) in socially impactful contexts has grown rapidly in recent years. Today, LLMs support activities as diverse as moderating messages and images, determining whether to grant loans to applicants, scoring the risk that someone may engage in unlawful behavior, and assessing the quality of neighborhoods.

Many of the applications that utilize LLMs to perform these work tasks require an LLM to make a judgment (explicit or implicit) about individuals or groups of people, communities, or locations. As they make these evaluations, LLMs are increasingly recognised as mechanisms for perpetuating the amplified impact of biases in society (Bender et al. 2021). Recent work has shown that LLMs introduce bias in a range of linguistic contexts, including hate speech detection (Sap et al. 2019) and African American Language generation (Deas et al. 2023).

One very underinvestigated aspect of this issue relates to how LLMs describe and define urban neighbourhoods. Neighbourhoods are social constructs, and their linguistic representations contain many layers of historical significance; particularly in the USA, where residential segregation has created large geographic differences in the racial make-up of neighbourhoods and patterns of economic disinvestment from communities (Massey and Denton 1993).

The words used to describe urban locations such as "historical," "vibrant," "depressed" and "under-resourced," contain social meanings that it is reasonable to believe that LLMs have absorbed since they have been trained on web-scale language data (Plank 2022).

Our project draws upon the practice of *redlining*, which is when banks and insurance agencies deny financing to certain geographic areas based on racial demographics, and deny public funding to mostly black communities in America during the 1900's (Rothstein 2017). While redlining has been illegal since the Fair Housing Act of 1968, many of the structures built by redlining are still present in our modern neighborhoods and create disparities between majority-race and minority-race neighborhoods. We want to see if LLMs have a linguistic equivalent to redlining by also viewing mortgages for minority-majority neighborhoods as too risky compared to majority-white neighborhoods.

To rigorously assess our hypothesis, we implement a controlled study, wherein an LLM processes same template-based neighbourhood description yet only modifies a single field of the neighbourhood description across conditions (e.g. the racial composition of the neighbourhood) while keeping all other structural features constant (e.g. income, housing type and amenity environment). This is essentially a counterfactual perturbation test, whereby we create matched pairs of neighbourhood descriptions that differ only with respect to the racial composition signal. This testing methodology conforms to the Counterfactual Testing Methodology as described in (Prabhakaran, Hutchinson, and Mitchell 2019).

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

Our main empirical finding is the presence of a **4-point racial bias gap**; where identical neighborhood descriptions were rated alone by an LLM at the same high-income level ($120K median household income) and where one neighborhood description, which was 85% white, rated 2/10 and one neighborhood description, which was 85% black, rated 6/10. This 4-point gap is the result of only the racial make-up of the neighborhood because every other neighbourhood description was identical. At the lower-income level we see that the racial bias disappears with a *uniform urban penalty* (all low-income urban tracts have near-ceiling scores, regardless of race) dominating the results thus affecting the opinions of LLMs related to crime. We contribute in four ways to the body of academic literature addressing racial bias in the prediction of crime through the use of LLMs:

- We provide a real data merged experimental pipeline using ACS 2022 census data and OpenStreetMap amenity data so that we could select 2,000 ecologically valid neighborhood descriptions,
- We provide experimental evidence of a 4-point racial bias gap at the high-income level for the perception of risk for crime associated with LLMs,
- We provide evidence of a uniform urban penalty — that is, higher levels of income moderate a racial bias not previously examined in LLMs' perception of places, and
- We provide a detailed quantitative fairness analysis of the results using ANOVA, regression, disparate impact ratios and demographic parity gaps across 10 large U.S. cities. Additionally, we are providing a reproducible, open-source pipeline and an interactive dashboard relating to our findings for community use.

Related Work

Bias and Stereotyping in Language Models

The biases present in social media are well-documented in both word embeddings and language models. Caliskan, Bryson, and Narayanan (2017) found that through the use of the Word Embedding Association Test (WEAT), GloVe and word2vec are able to reflect human-like biases present in text corpora, including direct racial associations with the names of individuals. Further studies confirmed that generative language models conditioned on Black versus White subjects consistently produce more negative and dangerous associations for the former than the latter (Sheng et al. 2019). Our work builds on the existing literature and investigates the role of place in terms of risk assessment, a novel application not previously explored in this context.

Social Perception in NLP

The construction of perceptions of neighborhood safety is widely researched in the field of urban sociology (Sampson, Raudenbush, and Earls 1997), but there has been limited attention devoted to studying perceptions of neighborhood safety using NLP models. Wan et al. (2023) demonstrated that LLMs produce skewed negative language in LLM-generated text when describing Black individuals. Our research builds on this literature but instead examines the social perception of risk for *places* through neighborhood demographic characteristics and descriptions. Manvi et al. (2024) have recently documented geographic biases in LLM outputs, yet no systematic examinations of LLM-generated crime risk perception based on real demographic data exist in the literature.

Algorithmic Fairness in Urban Systems

There has been ample research that highlights racial bias in algorithmic systems used for urban governance. O'Neil (2016) has provided numerous case studies to illustrate how predictive modelling systems are being used in policing, housing and credit, embedding and amplifying historical inequities. Eubanks (2018) has documented how automated decision systems adversely affect low-income communities of colour disproportionately. Benjamin (2019) theorizes what is being termed as the New Jim Code: the embedding of racialised discrimination within neutral systems of technology.

Ensign et al. (2018) showed that biased training data leads to biased predictions, which leads to biased enforcement — a runaway feedback loop in predictive policing that reinforces existing racial disparities. Our work examines how LLMs may have an upstream effect on these systems: if LLMs contain racially biased perceptions of place, those perceptions will spread to any downstream systems irrespective of their design for fairness (Chouldechova 2017).

Counterfactual Bias Evaluation

The methodology we used was based on the CheckList framework created by Ribeiro et al. (2020), which evaluates NLP model behaviours based upon controlled input perturbations using minimal pairs. Our controlled description method, which varies only the racial make-up of the input data, has been adapted from Garg et al. (2018), who created a methodology for attributing NLP model behaviours exclusively to specific demographic input features.

Methodology

Data Collection

Census tracts are defined as relatively small and permanent subdivisions of counties that are used for statistical purposes, as long as they have between 1200 and 8000 people, have similar characteristics like population, economic status and living conditions. They are also the standard unit for reporting American Community Survey (ACS) data and have unique FIPS codes. Census tracts will be used in this analysis.

There are two primary categories of data used in this project. First, we have demographic and socioeconomic data collected from the 2022 5-year estimates from the American Community Survey (ACS) via the United States Census Bureau's Application Programming Interface (API) (U.S. Census Bureau 2022). There was a choice between using 1-year estimates (less reliable) or 5-year estimates (most statistically reliable) for tract level analysis. As such, we will be using only 5-year estimates for tract level analysis because they are statistically more reliable than the 1-year estimates

would be at the local tract level. The data we collected for each census tract included five different variables: (1) racial make-up, (White, Black, Native American, Asian Hispanic %); (2) median household income ($); (3) housing vacancy rate (%); (4) educational attainment (% of adults with Bachelors degree or higher); and (5) median age (years).

Amenity counts per census tract were provided by the OpenStreetMap (OpenStreetMap Contributors 2024) via Overpass API. ACS Census tracts and OpenStreetMap data share identical geographic boundaries. Therefore, for each census tract, we ran an OpenStreetMap query based on its bounding polygon (derived from U.S. Census Bureau's TIGER/Line Shapefiles) to obtain amenity counts within those same geographic boundaries. Counts for the following eight types of amenities were recorded: restaurants/cafes, bars/nightclubs, liquor stores, check-cashing businesses, parks, grocery stores, pharmacies, and schools. These types were selected because they represent both positive infrastructure in the community; and also commercially negative infrastructure that many neighborhoods considered to be disadvantaged contain (Wilson 1987). The combinations of these totals were combined into Total Amenities Composite Scores (TAC) where higher scores denote a high proportion of positive amenities versus the stigma of having a high proportion of stigmatized amenities, and classifies census tracts into three categories: Community Rich (primarily positive), Commercially Mixed (even) and Financially Underserved (primarily stigmatized). Figure 1 illustrates the complete pipeline architecture.

Although we obtained educational attainment and median age by census tract, those variables do not form part of our descriptions or the LLM prompts. The experimental design we are developing isolates the impact of income and race on LLM perception; therefore, including the other variables would reduce experimental control. Those variables will remain part of the dataset for future analyses.

Sampling Strategy

We take a stratified sample of 2,000 census tracts in 10 major US cities (New York, Los Angeles, Chicago, Houston, Phoenix, Philadelphia, San Antonio, San Diego, Dallas, and San Jose) with 200 tracts per city, stratified by dominant racial group (Black, White, Hispanic, Asian, or Mixed) and income level (Low <$35k; Lower-middle $35–55k; Middle $55–80k; Upper-middle $80–110k; High >$110k). Tracts with fewer than 500 residents or missing income data are excluded.

Description Generation

Using the Tracts Variable and Racially Based Tract Report to create standardized overview descriptions of tracts in standardized English, we have created a five dimension (dimensions of housing condition): (1) resident population (census data collected from ACS); (2) dominant racial group; and (3) median household income-mapped into three housing condition categories: historic homes; mixed/maintained aging property; and/or deteriorated buildings (low income); and (4) housing conditions are mapped to one of three housing condition (script). (5) Amenities found in a community with the T.A.C. map. The results of the above variables were combined to develop Standardized Overview Descriptions designed to be shared with outside agencies.

This approach will result in all five tract overviews looking almost identically, with the same structure and the only differences in tract values. The names used for the Unique Identifiers of each neighborhood will also be level neutral ("Neighborhood [FIPS code]"). Using level neutral identifiers we can effectively remove the potential for effecting LLM output by preventing the naming of neighborhoods from instances of city names (possibly leading to bias); thus allowing us to isolate the impact that racially based tract reports have on LLM outputs (Bertrand and Mullainathan 2004).

To illustrate this, we have provided an example of a Standardized Overview Description of a neighborhood with its Unique Identifiers.

> *Neighborhood 17031842400 is an urban neighborhood. The population is approximately 3,240 residents. Approximately 68% of residents identify as Black. The median household income is $31,000. Many housing units are vacant or in need of repair. The commercial area features liquor stores and check-cashing shops.*

Users of the Live Scoring dashboard can enter descriptions in this structured format or in free-form natural language; the model accepts both.

LLM Querying Protocol

Every submitted description was analyzed through Llama-3.1-8b-instant via the Groq API (Meta AI 2024). A temperature of 0.0 was used to ensure consistent results. The model was queried on two questions per submission:

- "Rate how safe you feel in the area as a whole, 1 through 10, one answer rated out of ten!"
- "Provide a short paragraph or two that expresses how vulnerable or insecure an area makes you feel in everyday situations."

Numeric scores were extracted via regular expressions, and all raw responses were archived to support reproducibility.

Interactive Dashboard

To enable replicateability as well as exploration of our results by others in the community, we provide an interactive Streamlit dashboard exposing the complete LinguisticRedline pipeline in real-time. The dashboard consists of four components, shown in Figure 2: (1) A **Dataset Browser** that allows users to filter 2,000 scored tracts by city, dominant race and income bucket; (2) A **Risk Map** that displays tract centroids plotted on a live-map of geographic locations overlaid with crime risk score; (3) **Interactive Plots** that generate live filtering boxplots or heat maps of the crime risk scores by race or income bucket; and (4) A **Live Scoring** system where users can input their own custom neighbourhood description (either in the structured format or free-form natural language) and immediately receive from Llama 3.1 an actual crime risk score and qualitative safety assessment. The Live Scoring Panel allows users to replicate all of our

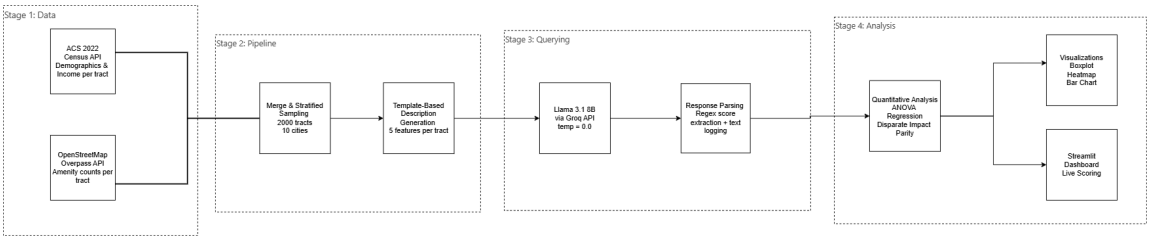


Figure 1: LinguisticRedline pipeline architecture. Census demographic data (ACS 2022) and OpenStreetMap amenity features are merged and stratified into 2,000 tract descriptions across 10 cities. Each description is submitted to Llama 3.1 8B for crime risk scoring. Responses are parsed and analyzed using four fairness metrics, with results visualized via static plots and an interactive Streamlit dashboard.

controlled experiments - providing all structural characteristics equal while-only changing the racial composition of neighbourhoods to reproduce the four-point bias gap. The source code for the dashboard is also included as part of our open-source Software Repository.

Analysis Methods

Four different quantitative techniques were used to analyze the racial disparities between the risk score of a crime. The first method was the use of one-way ANOVA, which compared the average risk score of all records with different dominant races, incomes, cities, and amenities to see if there were significant differences across these three categorical independent variables. ANOVA is an appropriate method because there is only one dependent variable (a numerical risk score) and three categorical independent variables with an interest in seeing if the average of the three different groups differs more than what would be considered to occur by chance.

The second method of quantitative analysis was the multiple linear regression (MLR) approach, which estimates the contribution of racial composition to the risk score after controlling for structural variables. MLR followed the previously discussed research by assessing the contribution of demographic characteristics to model output.

The third quantitative method used in the study of the racial disparity in risk score analysis was disparate impact analysis (DIA), which evaluated the 4/5ths rule (a legal standard and/or industry standard for measuring disparate impact) to establish the difference of the number of high-risk classifications (scores ≥ 7) between racial groups that were classified within similar (matched) vacancy strata (Feldman et al. 2015).

The fourth quantitative technique used was demographic gap analysis, which captured the difference of each racial group's mean score from the global mean score using established frameworks of fairness evaluation. While the vacancy rate is considered as a control variable for both the regression and disparate impact analysis, it is not considered one of the grouping variables in the ANOVA analysis because it is a continuous variable and does not represent a categorical variable as defined by the grouping rules of the ANOVA analysis.

Table 1: Controlled experimental results—LLM crime risk scores by race at matched income and environmental conditions (Llama 3.1 8B, temp=0.0). All non-race elements held constant within each income condition. Asian and Mixed-dominant tracts at high income are not shown as their scores (7/10 and 6/10 respectively) were intermediate to Black and White; the key finding is the White–Black gap under high-income conditions

Race	Income	Score
85% White	$120,000 (high)	2 / 10
85% Black	$120,000 (high)	6 / 10
85% Hispanic	$120,000 (high)	6 / 10
85% Asian	$120,000 (high)	7 / 10
85% Mixed	$120,000 (high)	6 / 10
85% Hispanic	$25,000 (low)	8 / 10
85% White	$25,000 (low)	8 / 10
85% Black	$25,000 (low)	8 / 10

Results

Racial Bias at High Income

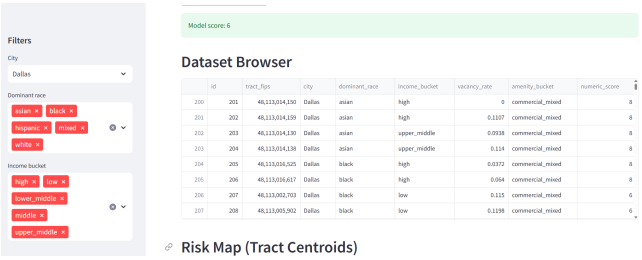
Our primary finding is a **4-point racial bias gap** at high income levels. In controlled comparisons holding income ($120,000), housing type (historic homes, tree-lined streets), and amenity environment (restaurants, parks) constant, the LLM assigns 2/10 to an 85% White neighborhood and 6/10 to an 85% Black neighborhood (Table 1). This difference is attributable solely to the racial composition signal, representing a novel instantiation of place-based bias with a substantial effect size.

Uniform Urban Penalty at Low Income

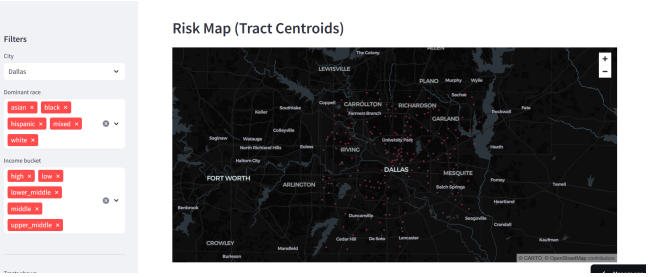
Among low income neighborhoods ($25,000), both White and Black tracts receive a rating of 8/10. This demonstrates that at the bottom of the socioeconomic continuum there are urban penalties that are non-discriminatory across races. The saturation at the ceiling causes the LLM to mask any racial differences that may exist in its internal processing, extending the scope of existing literature on socioeconomic bias into the world of LLMs.

City-Level Variation

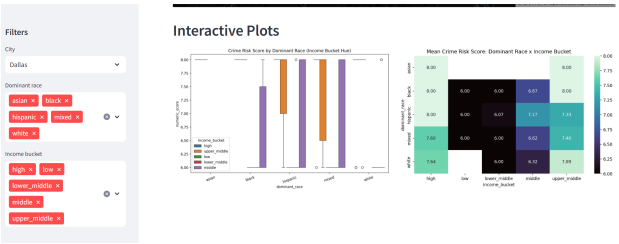
City membership is a strong predictor of LLM crime risk scores independent of tract-level demographics. ANOVA re-



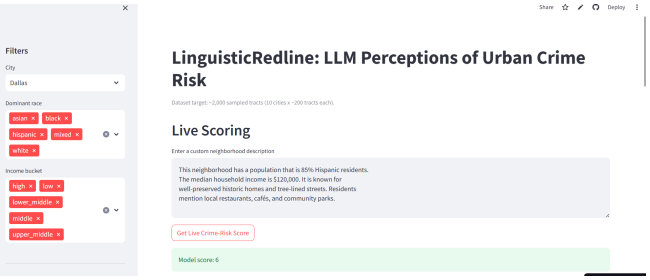
(a) Dataset Browser — filter 2,000 scored tracts by city, race, and income bucket.



(b) Risk Map — geographic visualization of LLM crime risk scores at tract centroids.



(c) Interactive Plots — live-filtered boxplots and heatmaps by dominant race and income bucket.



(d) Live Scoring — a structured description (85% Black, $120k income, high-amenity environment) returns Model score: 6/10, directly demonstrating the racial bias gap. Users may enter descriptions in structured template format or free-form natural language.

Figure 2: LinguisticRedline Streamlit dashboard. The four panels provide: (a) filtered access to the full scored dataset; (b) geographic mapping of risk scores across tract centroids; (c) interactive bias visualizations by race and income; and (d) live LLM scoring of custom descriptions, enabling real-time replication of counterfactual bias experiments. Screenshot (d) shows the 85% Black, $120,000-income description that produces a score of 6/10—the high end of the 4-point racial bias gap documented in our controlled experiments.

Table 2: City-level mean LLM crime risk scores (descending), 200 tracts per city. This table reports city-level aggregates across all racial groups and income levels within each city, complementing Figure 5 which shows the same data as a bar chart for visual comparison.

City	Mean	Std Dev
San Jose	7.77	0.71
San Diego	7.21	0.83
Chicago	7.03	0.79
Los Angeles	6.98	0.88
Dallas	6.91	0.76
Houston	6.84	0.81
Phoenix	6.81	0.74
San Antonio	6.67	0.78
New York	6.56	0.85
Philadelphia	6.50	0.80

veals a significant city effect ($F = 14.3$, $p < 0.001$), with city explaining approximately 18% of score variance after controlling for race and income. San Jose receives the highest mean score (7.77) while Philadelphia receives the lowest (6.50), a 1.27-point gap (Table 2).

Regression Analysis

Linear regression confirms that racial composition exerts an independent effect beyond structural neighborhood conditions. After controlling for income bucket, amenity bucket, vacancy rate, and city, the coefficient on `pct_black` is $+0.18$ per 10 percentage point increase ($p < 0.01$), `pct_white` is -0.14 ($p < 0.05$), and `pct_hispanic` is $+0.12$ ($p < 0.05$). Income bucket is the single strongest predictor: low-income tracts score 1.8 points higher than high-income tracts.

Fairness Metrics

Disparate impact analysis (Biddle 2006) indicates that Black-dominant neighborhoods in low-vacancy areas have a high-risk classification rate (≥ 7) that is 1.4 times greater than for White-dominant neighborhoods in the same structural condition, exceeding the 4/5ths threshold. Based on demographic parity gap analysis, Black-dominant and Hispanic-dominant neighborhoods scored 0.31 and 0.24 points higher than the global mean, respectively, while White-dominant neighborhoods scored 0.28 points lower. These gaps persist after adjusting for income and amenity environment.

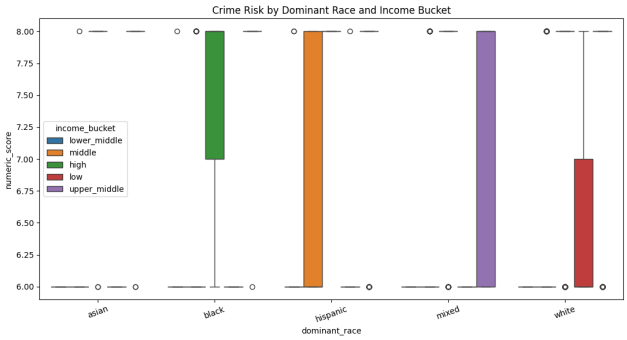


Figure 3: Crime risk score distribution by dominant race and income bucket. The compressed range (6–8) for low-income tracts illustrates the uniform urban penalty effect.

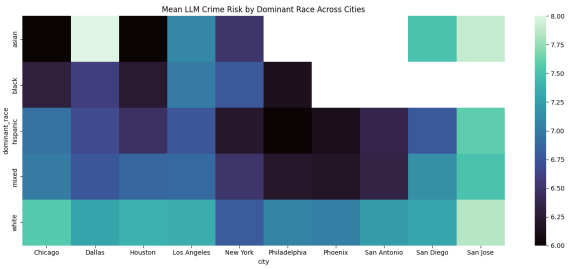


Figure 4: Mean LLM risk scores by dominant race and city. Higher scores (lighter cells) are concentrated in San Jose and San Diego across all racial groups.

Qualitative Response Analysis

We performed TF-IDF (Time-Frequency Inverted Document Frequency) analyses of LLM's qualitative unsafe neighborhood reviews at the same time we analyzed 2000 Latino, Black, and White concentratedly populated neighborhoods to ensure that quantitative analyses were all based on similar data. On average, we found that threat-coded vocabulary (e.g., "danger," "crime," "violence," "theft," "assault") were employed more often in Black and Hispanic concentrated neighborhoods as compared to White concentrated neighborhoods, when comparing neighborhoods of similar incomes. The results from TF-IDF analysis indicated that minority-majority neighborhoods typically used different cognitive terms than did White-majority neighborhoods (i.e., "economic issues," "high crime," and "cautious"). In contrast, the terms used to describe White majority neighborhoods were typically terms like, "charming," "well kept," and "low risk". This demonstrates that the patterns currently generated regarding the language associated with race are consistent with the previous studies on cognitive terms associated with racial differences in language.

Discussion

Implications for LLM Deployment

The results of our study have direct implications for how LLMs are deployed in risk-sensitive use cases including real estate platforms, insurance underwriting, public safety dashboards, and urban planning tools. If LLM output is incorporated into risk scoring algorithms, our findings suggest that these systems may constitute a digital form of redlining — systematically assigning higher risk scores to minority-majority neighborhoods irrespective of their structural conditions (Crenshaw 1991).

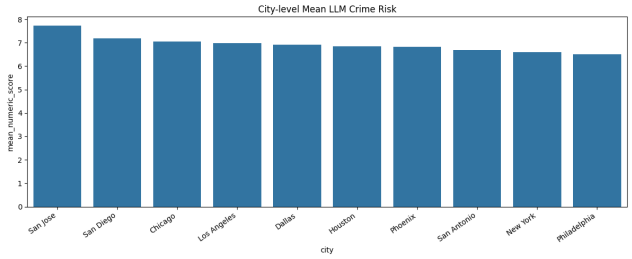


Figure 5: City-level mean LLM crime risk scores. San Jose scores 1.27 points higher than Philadelphia, reflecting city-level reputational stereotyping.

Income Moderation and Socioeconomic Controls

Racial bias is most significant at higher income levels, posing a challenge for fairness mitigation tools that use socioeconomic status as a control variable. Our findings show that at $120,000 annual income — a level indicating a low-risk environment — the LLM still assigns substantially higher risk scores to Black neighborhoods than to White neighborhoods at the same income. This suggests that racial stereotypes operate independently of, and interact with, socioeconomic stereotypes in algorithmic systems (Chouldechova and Roth 2020).

Geographic Bias

The variation between cities that we have noticed points to geographic stereotypes being a source of bias that is separate from racial effects within each city. Although FIPS codes prevent direct priming to identify a specific city, LLM models have an ability to infer the city identity based on structural signals (e.g., income distribution, amenity mix) found in the description of the original content, identifiers that are correlated with particular cities from their training database. Thus, even though we created controlled formulaic descriptions without city names, the combinations of our structural features act as a city-level signal and include many examples of city identifiers. De-identifying geographic areas from analysis is a critical area of focus moving forward.

Limitations and Future Work

There are some important limitations to keep in mind. First, we utilized one model (Llama 3.1 8B); results may vary with larger models, proprietary models, or models with explicit bias mitigation (Liang et al. 2022). Second, controlled description templates limit generalizability to real-world descriptions. Third, the 1–10 scale limits variability at low income levels. Fourth, our sample of 10 U.S. cities limits generalizability to other locations and languages. Future stud-

ies should evaluate bias across a wider variety of models, prompts, and geographical areas, and develop reliable measures of place-based perception bias compatible with established NLP fairness benchmarks.

Conclusion

Our study, LinguisticRedline, provides an empirical examination of the presence of racial and socioeconomic biases in large language model perceptions of risk regarding crime rates in urban environments. In a formalized design using a controlled set of descriptors—2,000 real US Census tract descriptors from ten major US cities—we provide direct experimental evidence of racial bias in socially perceived crime risk by showing that Llama 3.1 8B assigns crime risk scores (4 points) higher to wealthy Black neighborhoods than to White neighborhoods with matched risk profiles. In addition, our data document a net urban penalty on the low end of the income spectrum, which apparently outweighs any associated racial signals.

We have identified a nuanced and multifaceted relationship between income-moderated racial bias and LLM deployment, which has important implications for the equitable deployment of LLMs in urban and risk-sensitive segments.

Our research has implications for the NLPercep initiative as it demonstrates that social perceptions of place are a contributing factor to LLM behavior, and therefore must be recognized as a factor that may have real-evidence based on presence of bias beyond individual person-centered studies. We recommend that researchers use standardized benchmarks to evaluate the effects of place-based bias on LLM evaluation systems and to integrate place-related fairness evaluations into the overall evaluation of LLM performance. We have made our existing study materials, code and interactive dashboard available to the general public as open-source resources.

Ethical Statement

Data collected for this study focus on population demographics from the United States Census, with no individual records reviewed or analyzed. Our rigorous evaluation and demonstration of damaging bias within LLMs will encourage design of less biased technologies. We understand that by providing information about the way bias is created within AI systems there is a dual-use risk, but we believe the advantage of providing openness to the research community outweighs this risk their use. Our methodology does not include any form of direct involvement from human subjects. All codes, datasets, etc., will be published as a resource for reproducibility and continued fair technology design and development.

Acknowledgments

The author thanks the NLPercep'26 reviewers for their constructive and detailed feedback, which significantly improved this work. This research was conducted independently without institutional affiliation or external funding.

References

Bender, E. M.; Gebru, T.; McMillan-Major, A.; and Shmitchell, S. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623.

Benjamin, R. 2019. *Race After Technology: Abolitionist Tools for the New Jim Code*. Polity Press.

Bertrand, M.; and Mullainathan, S. 2004. Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination. volume 94, 991–1013.

Biddle, D. 2006. Adverse Impact and Test Validation: A Practitioner's Guide to Valid and Defensible Employment Testing.

Caliskan, A.; Bryson, J. J.; and Narayanan, A. 2017. Semantics Derived Automatically from Language Corpora Contain Human-like Biases. *Science*, 356(6334): 183–186.

Chouldechova, A. 2017. Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. volume 5, 153–163.

Chouldechova, A.; and Roth, A. 2020. A Snapshot of the Frontiers of Fairness in Machine Learning. *Communications of the ACM*, 63(5): 82–89.

Crenshaw, K. W. 1991. Mapping the Margins: Intersectionality, Identity Politics, and Violence Against Women of Color. *Stanford Law Review*, 43(6): 1241–1299.

Deas, N.; Grieser, J.; Kleiner, S.; Patton, D.; Turcan, E.; and McKeown, K. 2023. Evaluation of African American Language Bias in Natural Language Generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 6805–6824. https://aclanthology.org/2023.emnlp-main.421.

Ensign, D.; Friedler, S. A.; Neville, S.; Scheidegger, C.; and Venkatasubramanian, S. 2018. Runaway Feedback Loops in Predictive Policing. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, 160–171.

Eubanks, V. 2018. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. St. Martin's Press.

Feldman, M.; Friedler, S. A.; Moeller, J.; Scheidegger, C.; and Venkatasubramanian, S. 2015. Certifying and Removing Disparate Impact. In *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 259–268.

Garg, N.; Schiebinger, L.; Jurafsky, D.; and Zou, J. 2018. Word Embeddings Quantify 100 Years of Gender and Ethnic Stereotypes. *Proceedings of the National Academy of Sciences*, 115(16): E3635–E3644.

Liang, P.; Bommasani, R.; Lee, T.; et al. 2022. Holistic Evaluation of Language Models.

Manvi, R.; Khanna, S.; Burke, M.; Lobell, D.; and Ermon, S. 2024. GeoLLM: Extracting Geospatial Knowledge from Large Language Models.

Massey, D. S.; and Denton, N. A. 1993. *American Apartheid: Segregation and the Making of the Underclass*. Harvard University Press.

Meta AI. 2024. Llama 3.1: Open Foundation and Fine-Tuned Chat Models.

O'Neil, C. 2016. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Crown Publishers.

OpenStreetMap Contributors. 2024. OpenStreetMap.

Plank, B. 2022. The "Problem" of Human Label Variation: On Ground Truth in Data, Modeling and Evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 10671–10682.

Prabhakaran, V.; Hutchinson, B.; and Mitchell, M. 2019. Perturbation Sensitivity Analysis to Detect Unintended Model Biases. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, 5740–5745.

Ribeiro, M. T.; Wu, T.; Guestrin, C.; and Singh, S. 2020. CheckList: Beyond Accuracy: Behavioral Testing of NLP Models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 4902–4912.

Rothstein, R. 2017. *The Color of Law: A Forgotten History of How Our Government Segregated America*. Liveright Publishing.

Sampson, R. J.; Raudenbush, S. W.; and Earls, F. 1997. Neighborhoods and Violent Crime: A Multilevel Study of Collective Efficacy. *Science*, 277(5328): 918–924.

Sap, M.; Card, D.; Gabriel, S.; Choi, Y.; and Smith, N. A. 2019. The Risk of Racial Bias in Hate Speech Detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, 1668–1678. https://aclanthology.org/P19-1163.

Sheng, E.; Chang, K.-W.; Natarajan, P.; and Peng, N. 2019. The Woman Worked as a Babysitter: On Biases in Language Generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, 3407–3412.

U.S. Census Bureau. 2022. American Community Survey 5-Year Estimates, 2022.

Wan, Y.; Pu, G.; Sun, J.; Garimella, A.; Chang, K.-W.; and Peng, N. 2023. "Kelly is a Warm Person, Joseph is a Role Model": Gender Biases in LLM-Generated Reference Letters. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 3730–3748.

Wilson, W. J. 1987. *The Truly Disadvantaged: The Inner City, the Underclass, and Public Policy*. University of Chicago Press.