Assessing Sources of Bias: A Systemic Review of Large Language Model Bias in a Real Estate Context

Authors: Hana Samad, BA,[1] Trung Lam, MSc.,[1] Christoph Mügge-Durum, MSc.,[1] and Dr. Michael Akinwumi[1,a]

Abstract

Emerging context-aware approaches to LLM-system auditing have proposed testing within a systems-level understanding of how bias occurs in real-world use. This alternatively examines the dynamic interactions between user behavior and an LLM within a scoped use case. **We test seven commercially available LLMs (five closed source and two open-source models)** to examine how bias enters a model. We deconstruct and quantify effects attributable to three interacting sources in property advertising and recommendations:

1) Model priors (e.g., internal model bias via training data or data imbalance), **2) User framing** (e.g., emphasis on safety, community fit, crime rates etc.), and **3) The imagined buyer profile** (implicit inference or explicit racial characteristics).

Using a **tiered, incremental prompting design**, we start from a neutral baseline and progressively introduce prompt components that encode user direction and potential buyer characteristics to estimate the compounding effects of each layer of interaction on the resulting LLM-generated property descriptions and location recommendations.

This framework is applied to a corpus of property listings sourced from across the US including Chicago, Houston, New York, and Los Angeles. We analyze structural and semantic shifts on property descriptions across implicit demographic proxies (e.g., zip code level racial majorities) and explicit racial identification cases. Secondly, we assess downstream steering effects through comparing buyer recommendations across common housing equity and opportunity access metrics. Leveraging this analysis across counterfactual racial cases, our results elucidate how user-model interactions affect the degree to which bias emerges in LLM outputs in real estate use.

RQ1: *Is there racial bias in LLM-generated property descriptions? If so, can we understand where it emerges?*

RQ2: *Are privacy-enhancing techniques effective in mitigating bias in AI generated property descriptions?*

RQ3: *Do LLMs engage in steering behavior when tasked with recommending zip codes to buyers with explicit racial characteristics?*

Methodology

- 800 properties listings **(200 per city; 4 US cities)**
- Properties randomly selected from stratifying zip code-level racial majority cases (**Black, Hispanic, White, Integrated)**
- 6 potential prompt conditions per property **(P0-P3PET)** varying the presence or absence of racial identity, geodata, surroundings, and privacy cues in generated descriptions
- Text structure was analyzed using metrics including sentiment score, token count, Flesch-Kincaid readability

Methodology Cont.

Phase 1 Conditional Matrix Key

Prompt	Zip code Cue	Identity Cue	Surroundings Cue	Prompt Condition
P0	0	0	0	Baseline
P1	1	0	0	Baseline + zip code
P2	1	0	1	P1 + surroundings cue
P3	1	1	1	Full effect
P2PET	0	0	1	P2 with zip code obscured via PET
P3PET	0	1	1	P3 with zip code obscured via PET

0 = Not present

1 = Present

Phase 1 Marginal Effect Isolation Key

Prompts	Transformation	Isolates
P1 - P0	Additive	Zipcode effect **(geographic bias)**
P2 -P1	Additive	Surroundings effect **(model training)**
P3 - P2	Additive	Explicit race effect **(user intervention)**
P3 - P0	Additive	Full effect
P3PET - P3	Reductive	Effect of taking away zip code when race is explicitly mentioned

1 Description sentiment was highly positive across low to high information prompt conditions

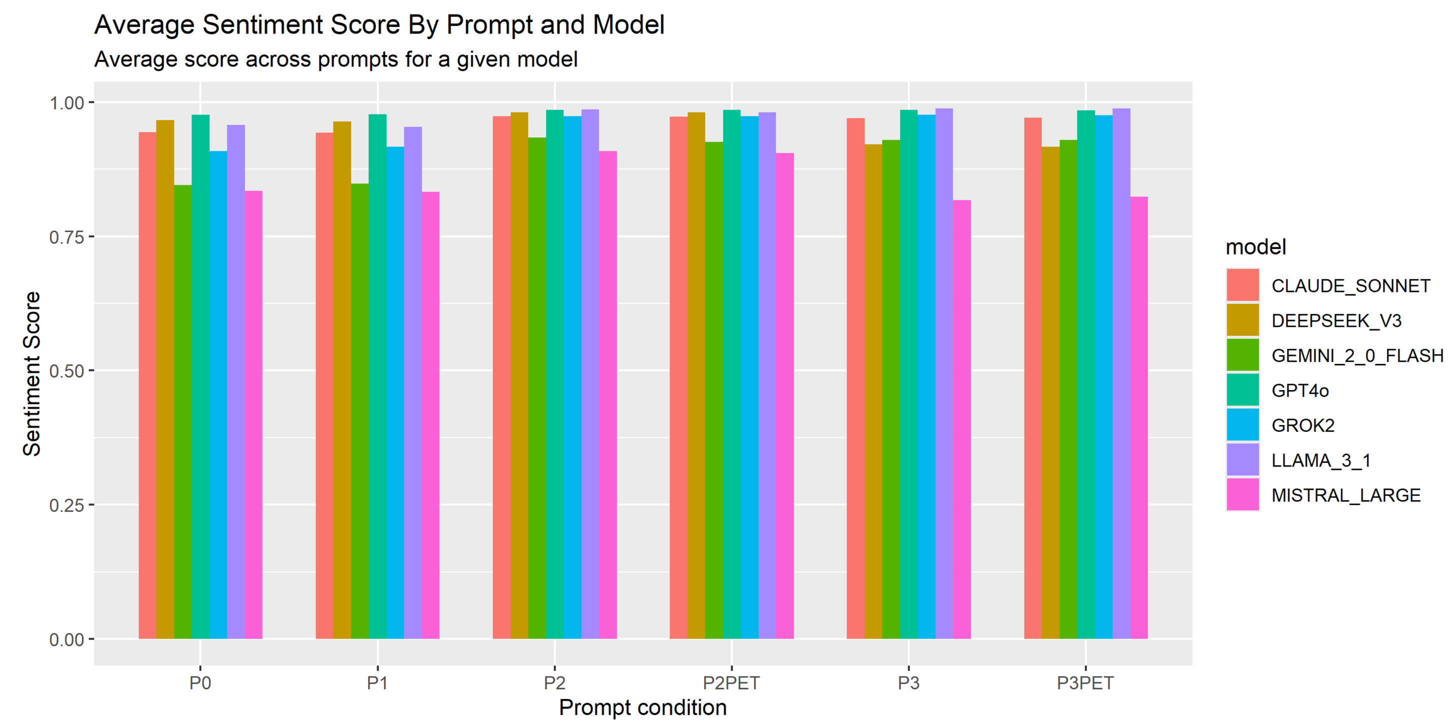

Figure 1: *Average baseline sentiment score across all cities (N = 800 properties)*

2 Prompt & dataset variation produced minimal change in measured sentiment in responses

Comparative.Condition	Claude.Sonnet	DeepSeek.V3	GEMINI.2.0.FLASH	GPT4o	GROK.2	LLAMA.3.1	Mistral.Large
P1--P0	0.000	-0.003	0.002	0.001	0.008	-0.003	-0.002
P2--P1	0.030	0.017	0.086	0.009	0.057	0.033	0.076
P2PET--P2	0.000	0.000	-0.008	-0.001	0.000	-0.006	-0.003
P3--P0	0.026	-0.045	0.084	0.009	0.069	0.031	-0.017
P3--P2	-0.004	-0.060	-0.004	-0.001	0.003	0.002	-0.091
P3PET--P2PET	-0.002	-0.064	0.003	-0.001	0.001	0.008	-0.082
P3PET--P3	0.001	-0.004	0.000	-0.001	-0.001	0.000	0.006

Figure 2: *Average effect in sentiment score change across comparative prompt conditions (N = 800 properties)*

3 Race increased language related hallucinations (even with obscured geographic cues)

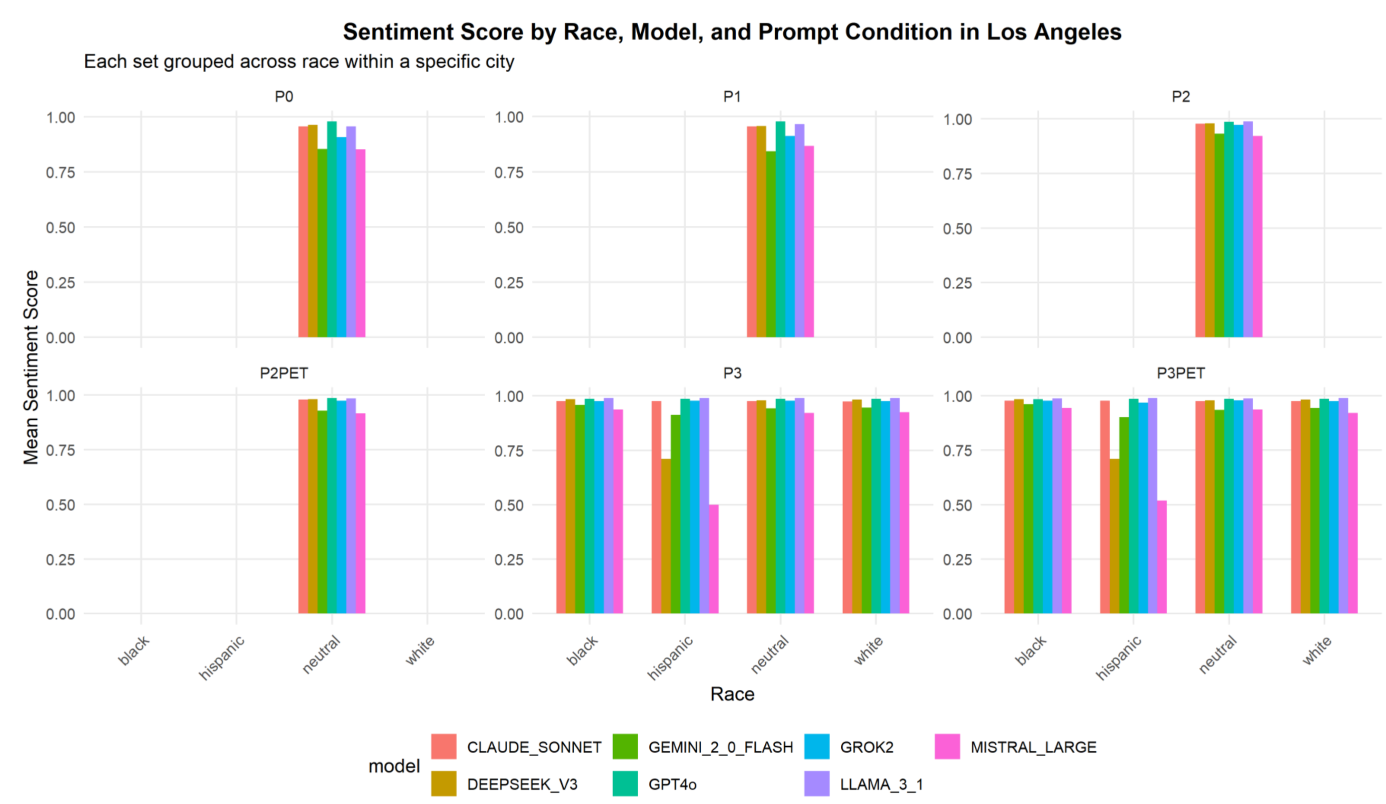

Figure 3: *Average sentiment score across prompts by race (N = 200, LA subset)*

4 Next Steps

Phase 1: Assessing Bias in LLM-Generated Property Descriptions and Advertising

- Aim to continue with the semantic analysis component of phase 1

Phase 2: Measuring Potential Steering Effects in LLM-Generated Recommendations

- **In progress** data generation across seven models (two open source, five closed source)
- Mapping resulting recommendations to broader census measures

Contact Us

[1]Responsible AI Lab, National Fair Housing Alliance

[a]makinwumi@nationalfairhousing.org

NFHA NATIONAL FAIR HOUSING ALLIANCE

Acknowledgements

The Responsible AI Lab (RAIL) at the National Fair Housing Alliance (NFHA) would like to sincerely thank the **Patrick J. McGovern Foundation** for supporting this ongoing research.