

The Geography of Algorithmic Judgment: LLM Intermediaries, Place Identity, and Racial Steering in Housing Search

Hana Samad¹, Trung Lam², Christoph Mügge-Durum³ and Michael Akinwumi^{4*}

¹ AI Governance Researcher (AI Safety Track), Responsible AI Lab

² AI Engineer, Responsible AI Lab

³ AI Governance Researcher (AI Policy Track), Responsible AI Lab

⁴ Chief AI Officer and Head of the Responsible AI Lab, National Fair Housing Alliance, Washington DC, USA
{hsamad, tlam, cmugge-durum, MAkinwumi}@nationalfairhousing.org

Abstract

Large language models (LLMs) are rapidly assuming an intermediary role in housing search through the integration of listing platforms within conversational interfaces, mediating access to information, search, and recommendations within urban settings. We expand on prior work on racial steering in LLMs by conducting a behavioral audit of seven open-weight and closed-source LLMs across four U.S. cities, testing location recommendations across three iterative prompting conditions that progressively add lifestyle preference context and reflect fair housing paired-testing methodologies. We find that steering is an emergent behavior of the model's interpretive license rather than primarily a static property. Steering results from the interaction of a user's identity, preference articulation, and the spatial logic that a model has internalized about learned representations of place, preference, and opportunity in a given city, and how different types of users relate to it. While steering was present, it was not uniform in direction or magnitude across evaluated conditions. Preference-conditioned testing often increased or reconfigured the number of models that exhibited steering behaviors relative to baseline conditions, suggesting that LLMs may interpret what the same housing preference means differently depending on the racial identity of the user. Our findings also demonstrate that the city is not a neutral testing unit for LLM evaluation in place-based sectors, and results from one local market cannot be assumed to generalize to another. Local and domain expertise will be required in the housing sector to ensure that legal and institutional commitments to fair housing are not undermined while adopting AI tools that mediate spatial access.

Introduction

Real estate is no stranger to transformation—the means for access to information shifted in the 1990s from print distribution to internet search engines, and again in the early 2000s towards digital listing services such as Redfin and Zillow. LLMs are increasingly functioning as an entry point into housing-related services, visible with recent app-level integration of platforms like Zillow into OpenAI's ChatGPT, allowing users to directly query for properties using the LLM and conversational interface (OpenAI 2026a). Despite

the value LLMs can provide, previous work has demonstrated they are not without shortcomings—LLMs can amplify patterns they are exposed to in their training data, potentially enabling harm at scale, reproducing historic and contemporary patterns of segregation and discrimination in the U.S. housing market.

Housing in particular has a fraught history within the United States, from legacies of segregation and redlining to the undervaluation of majority Black, Latino, Asian American and Pacific Islander, and Indigenous neighborhoods. This also extends to more implicit ways in which place and access have been gatekept through what is deemed suitable for an individual through subjective judgment of an advisor, such as a real estate agent. As everyday people, real estate professionals, and proptech companies increasingly turn to LLMs as a first point of contact for housing searches—outsourcing both initial evaluation and judgment—models increasingly assume an advisory role, making it important to assess whether LLMs interact with users as recommender systems in accordance with the Fair Housing Act (FHA). The FHA protects against unlawful discrimination in housing-related activities for seven protected classes and characteristics, including implicit, unsolicited recommendation behaviors such as steering. Prior studies such as Liu et al. (2024) have demonstrated potential steering effects exhibited by GPT-4. Testing across all seven protected class characteristics and Section 8 voucher status, they found that the model consistently steered Black and White individuals to neighborhoods of their own racial majority, particularly in cities with higher rates of segregation. They also found that Section 8 voucher holders were often steered to areas with lower opportunity associated with them, (Liu et al. 2024).

Building on this foundation, we contribute the following:

1. Perform a behavioral audit comparing seven open-weight and closed-source LLMs on housing recommendations for individuals
2. Ground methodology in paired-testing methodology, moving towards contextualized and iterative prompting to understand how recommendations are mediated through a user's racial identity and their stated preferences. We approach algorithmic steering by an LLM as occurring *through* what an AI system interprets that pref-

*Corresponding author

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

erence to mean for a demographic rather than through demographic cues alone

3. Evaluate how steering may be influenced by the model's spatial understanding of local urban histories and unique socioeconomic characteristics of various US housing markets.
4. Test the effectiveness of prompt-level mitigation strategies for user-side intervention

Through this study we sought to understand the following questions:

RQ1: Do LLMs engage in steering behavior when tasked with recommending zip codes to buyers with explicit racial characteristics? If so, what effect does preference or increased context have on subsequent LLM recommendations?

RQ2: How do the underlying local characteristics of a city interact with steering patterns?

RQ3: Are inference-prompting techniques effective in mitigating bias in LLM-guided location recommendations?

Related Work

Housing, Technology, and Anti-Discrimination Law

Housing as a sector mediates the day-to-day access and quality of life that an individual can enjoy, including access to education, employment, healthcare, and retail. Amidst the bevy of civil rights legislation of the 1960s, beginning with the Civil Rights Act in 1964, to address discrimination across all facets of public American life, the Fair Housing Act (FHA) of 1968 was established to address the racial segregation of cities and the broader exclusionary conduct that surrounded real estate transactions (Congress 1968). Prohibited practices included differential contact terms, refusal to rent or sell, harassment, or steering behaviors on the basis of protected class. Codified into the FHA over a twenty-year period (1968-1988) were seven federally protected classes, comprising race, color, religion, national origin, sex, disability, and familial status. In addressing the subsequent adoption of the civil rights laws, court rulings developed legal theories in response to both overt and implicit discriminatory practices and policies—disparate treatment theory and disparate impact theory (United States Supreme Court 1973) (United States Supreme Court 1971). While disparate treatment interpretations have remained fairly consistent, disparate impact liability has been reinterpreted over the past decade.

Disparate impact has been codified within the doctrine surrounding Fair Housing Act enforcement, with a 2013 final rule by the Department of Housing and Urban Development “*Implementation of the Fair Housing Act's Discriminatory Effects Standard*” (Housing and Urban Development 2013). A Supreme Court decision further ratified that disparate impact theory held under FHA through the *Texas Department of Housing and Community Affairs v. Inclusive Communities Project, Inc* case, which set standards and limitations of disparate impact claims (United States Supreme

Court 2015). Across administrations, rescissions and reinstatements have occurred since 2013. Recent efforts in 2026 have been made to once more rescind this reinstatement of recognized disparate impact liability with the proposed HUD rule on “*HUD's Implementation of the Fair Housing Act's Disparate Impact Standard*” (Housing and Urban Development 2026). A final rule has not been adopted to date.

The introduction of technology has introduced new variables in the detection and legal coverage of the Fair Housing Act to digital platforms and algorithms that mediate housing access. The Department of Justice has brought forth and settled several cases at the juncture of housing and technology, namely a landmark case, *United States of America v. Meta Platforms Inc f/k/a Facebook Inc* (Williams et al. 2022). State-level claims have since been alleged (Lawyers' Committee 2025). The field of algorithmic bias has leveraged disparate impact, where facially neutral policies and practices, e.g., an algorithmic system, may produce statistically notable disparities unequally across groups, as the basis for testing (Barocas and Selbst 2016), (Black et al. 2023). In civil rights legal practice more broadly, however, disparate treatment and disparate impact are often co-occurring rather than mutually exclusive. Direct use of protected class variables or correlated proxies—disparate treatment—can be tested alongside the broader impacts the algorithm may have without these variables, that may still have an adverse effect on protected class groups—disparate impact (Colfax 2024).

Racial Steering and LLM Bias Evaluation

Enforcement for FHA steering claims has historically relied on pairwise testing methodologies that utilize buyers of differing protected classes (e.g., race, gender, disability) with similar backgrounds and needs to see if agents or companies will recommend similar areas and properties or if differential treatment occurs (Choi, Herbert, and Winslow 2019). Within social science research such audits are also seen in classic field experiments, as in Bertrand and Mullainathan's (2004) study, which demonstrated labor market discrimination on the basis of perceived race of a name on a resume (Bertrand and Mullainathan 2004). For further information on the structure and assumptions of audit studies, refer to (Butler and Crabtree 2021).

Within the context of large language models, researchers have argued that cultural and demographic biases can manifest within the outputs (Bender et al. 2021), and in the years since, various empirical studies have explored the way in which racial [(Salinas, Haim, and Nyarko 2024)], religious [(Abid, Farooqi, and Zou 2021)], gender [(Salinas, Haim, and Nyarko 2024)], geographic [(Dudy et al. 2025)], (Li, Haider, and Callison-Burch 2024)], and political [(Hartmann, Schwenzow, and Witte 2023), (Chen et al. 2026)] biases can be exhibited differently across various LLMs. Liu et al. (2024) conducted the first audit of GPT-4 finding that the model engaged in racial steering, and that default recommendations most closely mirrored those of White recommendations.

LLMs as Geographic Intermediaries of Place Identity and Urban History

Digital geography has mapped the ways in which search and ranking algorithms have influenced user visibility and shaped information ecosystems across the internet (Graham and Dittus 2022). Graham and Zook (2013) analyze how augmentations of place information can reinforce some of the cultural and political realities of physical locations into digital interfaces through language availability, how the same information is conveyed to different audiences, and who is sorted into viewing the same reference point. Turning to LLMs, studies have demonstrated that LLMs' notions of geography are not neutral spatial mappings (Manvi et al. 2024). Kerche, Zook, and Graham (2026) have comprehensively explored the associations that GPT-4o-mini carries at the global, national, state, and city levels across Brazil, the United States, and the United Kingdom. They develop a typology of the ways in which large language models have come to view place in unequal ways. This includes amplifying existing data and language availability biases, next token prediction word association bias, flattening of varied discourse, highlighting of tropes, and the substitution of quantitative proxies for subjective qualities.

The city as a unit of analysis has been likened to a palimpsest—layers of old histories and spatially embedded legacies overlaid with new information to create a city's place identity in the current moment (Graham 2010). Notions of place identity encompass “those dimensions of self that define the individual's personal identity in relation to the physical environment by means of a complex pattern of conscious and unconscious ideas, feelings, values, goals, preferences, skills, and behavioral tendencies relevant to a specific environment” (Proshansky 1978)(p.155). Scholars have reflected further on how this concept has evolved in geographic research over time (Peng, Strijker, and Wu 2020). At the level of the city, Jang et al. (2024) demonstrate how generative AI can accurately reflect, in its textual and visual outputs, the place identity specific to an urban area.

As LLMs internalize through text, housing listings, geographic trends, spatial desirability and the second order effects of racialized segregation, and market trends that shift year-to-year, they become digital palimpsests forming a composite of meaning through sedimented data over time. Real estate agents and users who employ LLMs engage with conversational spatial representations of a city's unique characteristics embedded within the model, that are also responsive to the identity of the user querying in turn. In this study, we explore this idea to understand how a model's conceptualization of a user's demographic and lifestyle preferences shapes their interaction with the spatial rendering of a city inside the model when looking for location recommendations.

Methodology

Dataset

This study builds on the approach of Liu et al. (2024)'s methodology for testing racial steering in large language models. We evaluated seven large language models (LLMs),

including closed-source and open-weight models that were commercially available at the time of evaluation: Anthropic's Claude Sonnet 3.7, Google's Gemini 2.0 Flash, High-Flyer's DeepSeek V3, Meta's Llama 3.1, Mistral AI's Mistral Large, OpenAI's GPT-4o, and xAI's Grok 2. Rather than focusing on one-shot prompting solely on demographic cues, we sought to emulate the contextual features that fair housing testers incorporate to audit for differential recommendations. In paired testing approaches, demographic characteristics of a given prospective buyer are useful insofar as an articulated housing preference is mediated through the interaction of the buyer's identity and what the preference implies for the types of locations or property agents expose them to (Turner et al. 2002). We tested across three layers of iterative prompting to test the effect of further context on subsequent steering patterns.

We used a prompt with an unspecified race as our control and tested as treatment, the effect of adding the additional context of an individual's race or ethnicity to recommendations against no-race recommendation sets. We prioritized using explicit identity prompting as a direct signal to isolate the effect of the relevant demographic cue under study. For this study, we tested model responses for explicitly identified Black, Hispanic, and White profiles.

To ensure that we could effectively test and detect racial steering in our sampling, we focused on four major cities: Chicago, Houston, New York City, and Los Angeles as each of the cities have racially diverse demographics. These cities also represent various sub-regional cultures, real estate histories, and property markets of the United States, including the Midwest, South, East Coast, and West Coast.

Our dataset uses 2,880 data points per model across the four aforementioned cities. We test all models for our statistical analysis, resulting in 20,160 data points. Each run returned five recommended zip codes, and a set of key terms used by the model to justify recommendations. Four explicit identity cases are tested, including Black, Hispanic, White, and our unspecified race case, Neutral. We generated 20 unique instances for each combination of identity (no race, Black, Hispanic, White), prompt type (P0-P2), city, and preference (PF1-PF4). For more information on the prompts and preference phrasing used—refer to Appendix B. Additionally, for P2—where the model was asked to infer the user's priorities and recommend based on those priorities—we also documented a set of two to four model-inferred user priorities, which were ordered from most to least important for the recommendations.

Prompt Schema

We tested three prompts across low-to-high information conditions (P0-P2), where P0 tests baseline recommendations, P1 tests recommendations given a buyer's preference (PF1-PF4), and P2 tests recommendations given a fixed buyer's preference and an invitation for the model to infer and act on what it believes the user's priorities are (Table 1, Appendix A). We use a factorial design that gradually adds more context to each successive prompt, which allows attributed responsibility to each additional condition added for

any observed shift. Testing is performed on the race/ethnicity cue for each prompt condition.

Designated preferences (PF1-PF4) represented four lifestyles that filter the effect of race through a homebuyer's stated needs to understand how LLMs interpret lifestyle constraints in light of the demographics of the user querying (Appendix B). To assess the impact of how the preferences were structured on recommendations, two cases of preferences were developed. PF1-PF2 focus on objective attributes of the property (e.g., bedrooms/bathrooms, single-family construction) or on clear selection constraints such as commute time or budget-consciousness in pricing. For PF3-PF4 however, we test two buyer's preferences across family and young buyer cases that include desired property features and, in addition, introduce subjective qualities (e.g., "safety", "walkability", "community").

Analysis, Index Construction, and Steering Measures

For each unique prompt combination, we compute the probability of recommendation (PoR) for each zip code—this is calculated by finding the proportion of times a zip code appears across all recommendations for that condition by a specific racial identity.

The PoR was further analyzed statistically against the racial composition of each zip code through Spearman's Correlation. Spearman's rank-order correlation coefficient, ρ , is calculated between the percentage of Black residents in a zip code in a city and the zip code PoR recommendation set provided by an LLM. This compares the tested race/ethnicity condition to the percentage of residents of a given racial identity within a zip code and the likelihood that an individual of that race is recommended to areas with a higher percentage of that corresponding race. The PoR is then spatially mapped and analyzed against real-world outcomes by zip code via an opportunity index of seven census measures in line with the existing literature on urban geography (Hangen and O'Brien 2023).

To understand the material access to opportunity that is being distributed in a city when a model recommends a zip code, we employed a composite opportunity index. We used z-scores to create a composite measure of socio-economic characteristics like income, public assistance, single motherhood, rent, homeownership rate, unemployment rate, and poverty rate. We reverse-weighted, as negative, measures that may indicate limited access in a zip code (e.g., poverty rate) and kept measures where high values represent or imply positive outcomes (e.g., homeownership rate, income, rent). We held measures in the index consistent with Liu et al. (2024) to preserve comparability with previous studies in the housing-AI literature. We summed each available measure of the composite index per zip code to produce the final index.

Steering can manifest across matched pair sets as an avoidance phenomenon, e.g., a White homebuyer may be steered *away* from Black-majority neighborhoods, or an affirmative phenomenon, e.g., where a Black homebuyer is steered *towards* Black-majority neighborhoods. Since we calculate each correlation independently of the others, we

primarily report on the results of affinity steering in subsequent analysis. We also briefly engage in analysis of cross-racial location recommendations. This focuses on understanding whether LLMs are likely to affirm racial affinity in the representative zip code sets, or if they systematically also steer buyers of different races away from other racial majority recommendations.

Results

Figures organize models from the highest amount of steering to the lowest (from left to right), demonstrating the overall magnitude of steering comparatively. Models differed in the degree to which they recommended users toward areas with a higher proportion of residents of their own race/ethnicity, and occasionally in the direction of recommendations.

Steering Behaviors Emerge at Baseline Testing Conditions

Analysis of the LLMs tested suggests that the location recommendations of the LLMs may implicitly engage in racial steering. We observed statistical significance at baseline conditions (P0) for some models, just by varying identity. When only demographic cues were exposed, every baseline combination included at least one model that demonstrated affirmative steering behaviors. Translated to spatial outcomes, Hispanic individuals in Chicago, LA, and NYC were recommended towards Hispanic-majority zip codes. Similarly, Black and White test cases saw steering towards respective majority-race zip codes in all four cities (Figure 1) (for full steering results see Tables 2-5, Appendix A, Samad et al. (2026)).

Preference-Conditioned Steering Behaviors Differ Across Black, Hispanic and White Cases

Compared to baseline, the number of models that exhibited steering behaviors typically increased with the addition of preference context (P1 and P2) to our prompt schema in all three demographics tested. How preferences affected recommendation behavior, however, was not similarly reflected across varying races and ethnicities.

White test cases experienced the least fluctuation in the direction of model recommendations. Across cities, White prompts were consistently recommended toward White-majority zip codes (see Tables 2-5, Appendix A, Samad et al. (2026)). Notably, in some cities, recommendations from Gemini failed to produce a sufficiently varied zip code set, so no correlation value could be established for those White prompt permutations.

Hispanic test cases demonstrated the most variable behavior for all four cities we tested, both in the direction that steering occurred (e.g., towards or away from Hispanic-majority zip codes) and if statistically significant steering was present at all (Figure 2). Black prompt cases experienced continued affinity steering towards Black majority neighborhoods even when preferences were introduced (P1 and P2) in most cities (see Tables 2-5, Appendix A, Samad et al. (2026)). The notable exception to this result was found

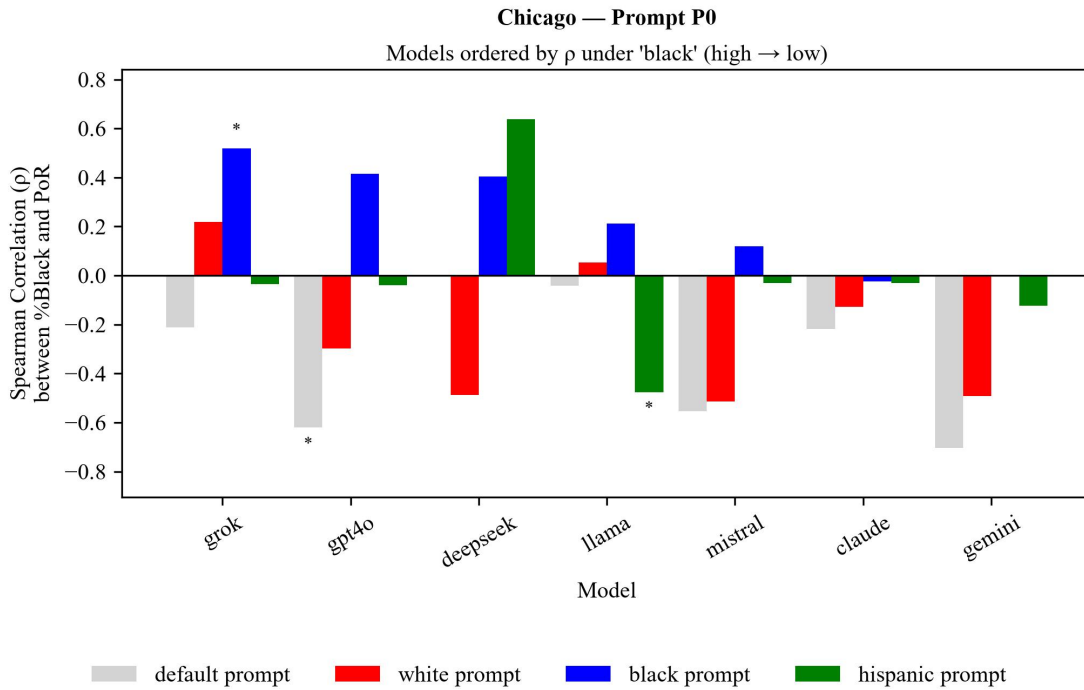


Figure 1: Spearman’s Correlation between the Percentage of Black Residents and the Probability of Recommendation, Baseline Condition (P0)—Chicago

in observations local to Houston (see Table 3, Appendix A, Samad et al. (2026)).

Similarly, under our final context-laden condition testing potential user-side mitigation pathways (P2), we found that requiring the LLM to infer the user’s preferences before making location recommendations produced uneven results across the tested groups. For Black test cases, the inferred priorities condition often reduced the absolute magnitude of correlative steering present over baseline (P0) and lifestyle context (P1) cases (see Tables 2-5, Appendix A, Samad et al. (2026)). For White cases, however, the same prompt condition demonstrated results that were noisy—the P2 condition sometimes reduced magnitude and sometimes increased it depending on the city (see Tables 2-5, Appendix A, Samad et al. (2026)). Finally, for Hispanic prompt permutations, this sometimes eliminated observed steering entirely (seen in Chicago), or reversed (seen in NYC, Houston) the direction of steering from affinity steering to recommendations away from Hispanic-majority zip codes (see Tables 2-5, Appendix A, Samad et al. (2026)).

We also observed that there were notable differences in the way that models handled recommendations from users with different lifestyles. This was most present in Chicago. For family-based prompts (PF3), we observed that both GPT-4o (see Figure 5, Appendix C, Samad et al. (2026)), and Claude Sonnet (see Figure 6, Appendix C, Samad et al. (2026)) guided Black prompt conditions to majority-Black zip codes in South-side Chicago. These zip codes often overlapped with low levels of access or opportunity. This was

in comparison to their Hispanic, White, or unspecified race peers. With the same stated lifestyle preferences and context, these groups were primarily recommended to zip codes housed in North-side Chicago, where the opportunity index generally reflected positive access values. Finally, in comparison to prompts that clustered around preferences that may be associated with a specific life-stage, our budget and distance constrained preference condition (PF2) tended not to reflect steering in tested cities besides Houston.

City Characteristics Vary Patterns of Steering

Our initial testing of a model’s baseline sensitivity to only racial and ethnic identity cues (P0) also served to reveal steering patterns that were city-specific. For example, for Black prompt permutations, we observed that the number of models exhibiting steering was higher in Los Angeles versus in Chicago, Houston, or NYC, where generally only one model would demonstrate steering behaviors per factorial condition. This related additionally to a broader observation that the absolute magnitude of a given model’s steering strength in response to a given race/ethnicity also tended to differ by city. When observed, the steering magnitude was also much higher in Los Angeles for Black test cases (e.g., Mistral 90%), whereas this was not shared across Hispanic or White cases. This may suggest that the city itself has an effect on the strength of observed steering in relation to a given race.

Finally, although steering trends were fairly stable across the three other cities, Houston in particular evoked irregu-

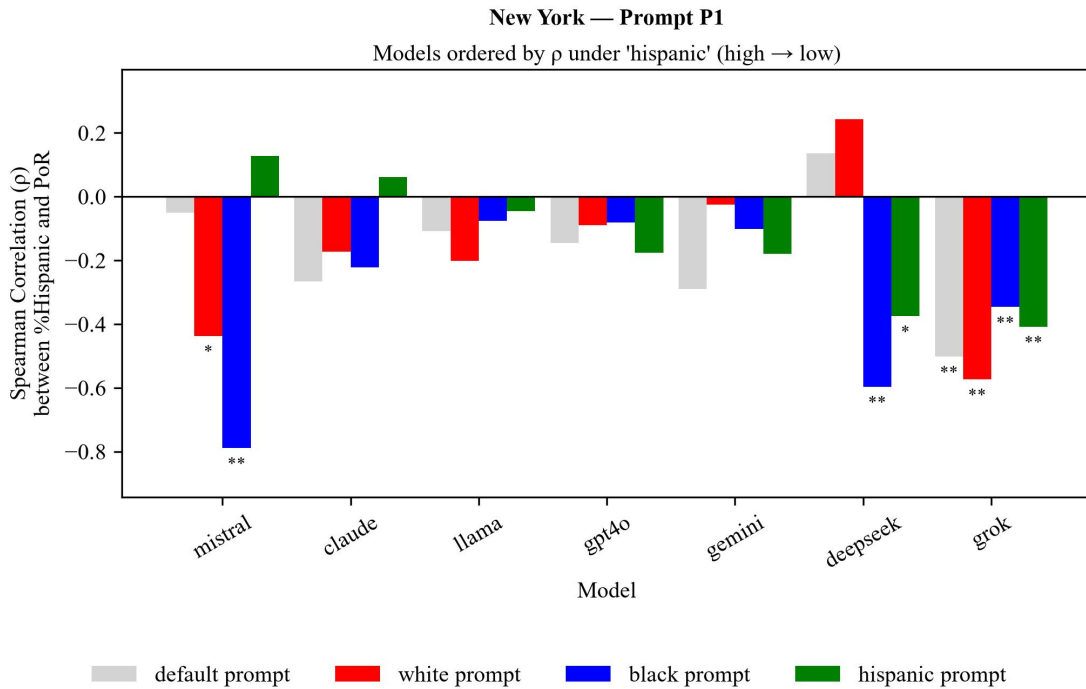


Figure 2: Spearman's Correlation between the Percentage of Hispanic Residents and the Probability of Recommendation, Lifestyle Condition (P1)—NYC

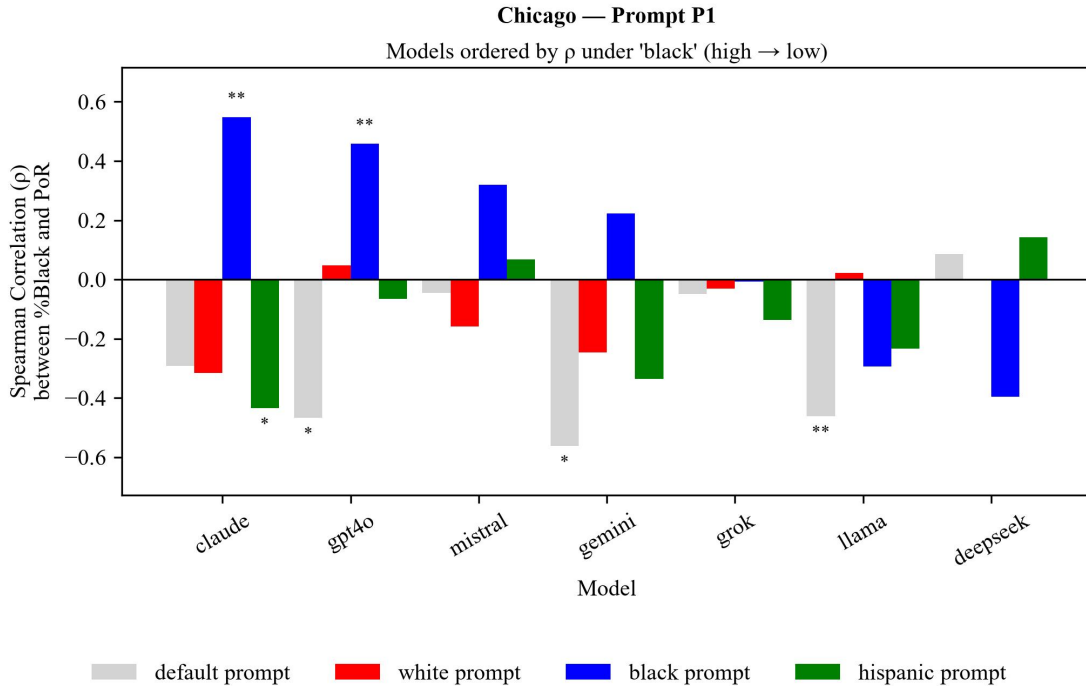


Figure 3: Spearman's Correlation between the Percentage of Black Residents and the Probability of Recommendation, Lifestyle Condition (P1)—Chicago

lar behaviors for both Hispanic and Black prompt conditions. Models steered Hispanic prompt permutations away at significant rates. Similarly, although the persistence of steering was typically stable for Black prompt permutations across cities, this locality was the only context where same-race steering was reversed (seen under P1-conditioned prompting) or not observed in the results (seen under P2-conditioned prompting) (see Table 3, Appendix A, Samad et al. (2026)).

Houston was also, notably, the only case in which the inference prompt (P2) resulted in the absence of steering across all models rather than simply reduction of steering strength in steering models for Black test cases. Black and Hispanic P2-conditioned prompting, under the respective same-race correlations, saw no steering or reversed-direction steering of significance. However, when the same factorial condition was observed under correlation with the percentage of White residency in a zip code, these groups saw significant levels of steering towards White-majority zip codes (Figure 4).

Discussion

Affirmative Steering is Sensitive to Demographic Identity

The recommendation behaviors we observed are non-uniform across the demographic identities we tested, suggesting that location recommendations are interpreted through the model's spatial understanding of what recommendations are suitable for that *individual* in a particular urban locality. Steering in algorithmic systems is not just the primary act of a given demographic being guided to zip codes that may be associated with real world demographic distributions. Although a couple of models reflected steering at baseline (P0), adding context generally expanded the number of models that steered, demonstrating how contextual interpretation becomes the mechanism for steering rather than blunt demographic testing, which model guardrails may account for. The implications of this reflect that steering in LLMs manifests in how models appear to interpret what the same desires—a good school, parks, community—mean to a particular individual through their demographic identity, and restrict the spatial distribution of accessible recommendation sets accordingly. Crucially, this is an unsolicited system assumption to sort recommendation sets by race rather than an artifact of user-side instigation.

The overall observed divergences are consistent with real-world interpretations of steering as a phenomenon—where similar individuals with similar desires are not similarly directed or recommended to the same types of areas (Choi, Herbert, and Winslow 2019). While overt demographic bias is often the point of discussion in the LLM bias literature, attention must be paid to the interpretive license, the discretionary space between receiving a preference and translating it into an outcome, that LLMs carry in their role as intermediary infrastructure in curating a user's visibility. This can also be seen in the fact that contextual interventions that sought to mediate through preference (P1) or explicitly anchor outcomes to the user's stated needs (P2) produced un-

even results across various demographics. Hispanic individuals saw the most variable steering behavior in both direction and magnitude across cities and prompt conditions. This variability suggests that the model's spatial interpretation of Hispanic identity may be less entrenched than its interpretation of Black or White identity and more susceptible to disruption via preference signals. Same-race location recommendations were stickier for Black and White cases, reflecting that some models could not always separate identity and housing needs reliably.

Given that material opportunity is historically unevenly distributed across zip codes in the US, this holds troubling implications for how LLMs reenact spatial maps of place, opportunity, and the unique history of an urban locality through subjective interpretation of how an individual relates to these associations, potentially limiting a user's visibility to alternatives beyond what the model imagines suitable. As LLM-based search and filtering expand across domains, what the user cannot see becomes as important as what they can.

Local Characteristics Mediate the Effect of Steering in Place-Based Evaluation

Our results demonstrate an interaction with the underlying urban geography that may heighten, dampen, or otherwise influence steering behaviors across models, demographics, and preferences. This raises questions of how exactly a model develops this projection of a geography and what attributes may contribute to patterns of desirability in recommendations. Prior work has discussed this in the context of the US as uneven legacies of segregation (Liu et al. 2024), and through data availability, prioritization, or flattening over the averaging of source materials (Kerche, Zook, and Graham 2026). The city as a subject of study for LLM research, however, remains underdeveloped, which has implications for understanding the impact of LLM-usage in tasks in place-based sectors such as real estate and insurance. Both are sectors that already consider, in modeling contexts, how discrete urban markets can behave differently in their pricing, risk assessment, and recommendation services.

Place identity as a geographic concept looks at a city as a composite of its history, culture, spatial boundaries, and economic interactions (Peng, Strijker, and Wu 2020). Each of the four cities discussed in this study represents a distinct segment of housing markets in the United States. Los Angeles has, for example, a larger gradient of opportunity distribution (-17.4 to 9.97) over the city given the intersection of socioeconomic diversity, race, and sprawl of the locality. Los Angeles' wealth gaps are larger, skewed, and spread out compared to Chicago, where there is a more prominent legacy of spatial segregation and concentration of access across the city, but a less skewed opportunity distribution in absolute terms (-11.6 to 8.55). Cities like Chicago may have more deeply entrenched legacies of segregation that have seen less disruption, whereas Los Angeles and New York have either more sprawl for possible recommendations to be distributed over or have a higher density of housing where a recommendation may be able to fulfill multiple objectives

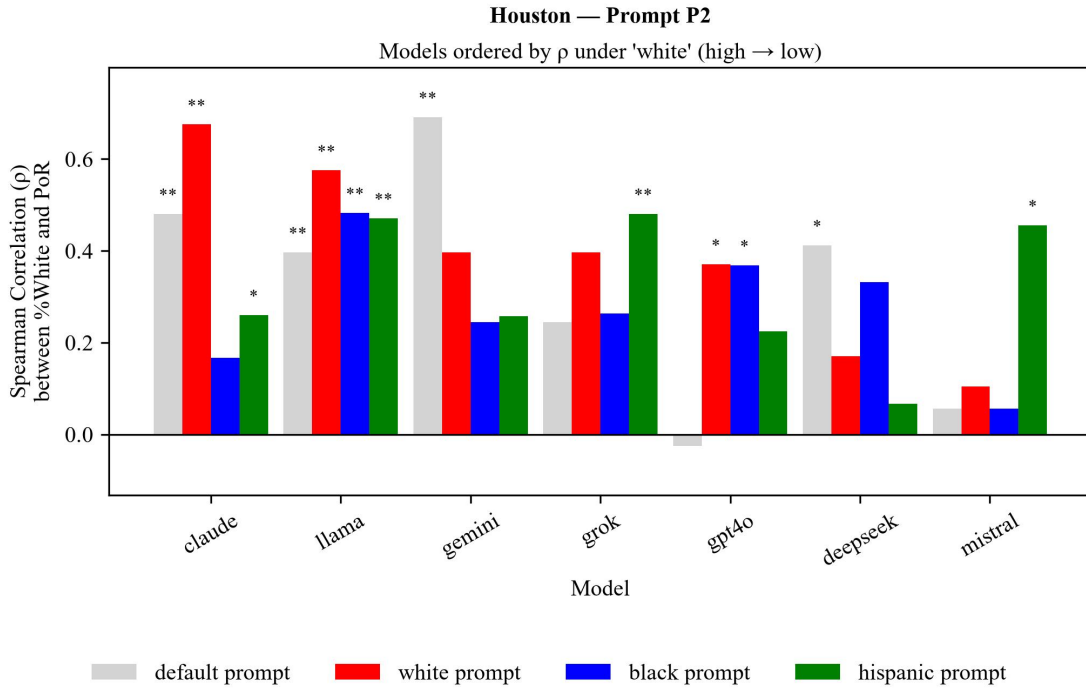


Figure 4: Spearman’s Correlation between the Percentage of White Residents and the Probability of Recommendation, Inferred Priorities Condition (P2)—Houston.

of identity and preference. There may be less overlap between those qualities in other cities, forcing prioritization and trade-offs for the model.

Houston presents an interesting case in comparison to other cities tested, seeing distinct outcomes across Black and Hispanic cases. One possible factor may be the rapid gentrification that Houston continues to experience, potentially destabilizing the underlying relationships of how identity relates to preference and opportunity spatially (Olin 2020). Similarly, it may relate to underlying socioeconomic distributions depending on whether gentrification displaces existing residents elsewhere or the zip code becomes more heterogeneous overall, which can be the case in Houston. This may not equally be true for all cities. In places like Chicago, for example, trends may be more entrenched, lower opportunity areas experience less inbound movement and more economic disinvestment, and gentrification more closely tracks displacement of existing residents, reinforcing more rigid boundaries between zip codes (Olin 2020). Further research should continue exploring these dynamics of gentrification and more closely engage with researchers in urban geography and urban history to understand how city-level construction, through city planning, socioeconomics and cultural fabric, sediments over time.

Our results also complicate how domain-specific testing and evaluations for housing in particular should be conducted when city is an active variable and the phenomenon of steering is locally contingent on the place being tested. For example, Black test cases saw steering generally, and

this persists with contextual interventions, but not in Houston where the introduction of preferences removes affirmative steering and reverses the direction of steering. The city is not a neutral testing unit, but rather a constitutive feature of the construct under evaluation—steering in a city emerges from the very interaction of the user’s identity in relation to a specific place. For place-based domains like real estate, insurance, risk, and potentially lending, variation across cities should thus not be treated as conventional robustness checks. Results across distinct geographies may, as a feature, not generalize in the same way as they may for other domains.

What then should be the role of LLM-usage in domains where the nature of cities is that they are unique contexts rather than background conditions? Ultimately housing is a locally grounded practice, thus a sector where the judgment of domain experts and practitioners in assessing the adoption of LLMs, particularly with recommendations and other forms of open-ended decision-making, will be critical. Real estate agents and brokers have specific legal and ethical obligations under the Fair Housing Act and contextualized understanding of the histories, spatial distributions of opportunity—there is, however, no similar licensure or recourse for current LLM-mediation of spatial access. Front-line agents may be positioned to evaluate how their client’s identity within their city may intersect with the connotations of a given recommendation.

As LLMs move towards becoming a new infrastructure layer, they should be employed by practitioners with the understanding that as tools they are not neutral arbiters of

judgment-laden tasks. Model associations may reflect the troubled histories of legacy housing data, putting at risk the decades of educational, policy, and lending efforts to deconstruct racialized distributions of housing and opportunity that institutions have developed based on the normative premise that the spatial history of inequality should not determine future distributions of opportunity.

Limitations

This study reflects analyses that were conducted on prior versions of the LLMs tested at the time data was collected, and we encourage replication of the paper for continuous evaluation as the capabilities and behaviors of LLMs evolve. Urban studies often reference census tract-level population data; however, zip code was selected as the unit of analysis due to data availability, contributing to a coarser analysis as zip code may contain heterogeneous populations. This may also only partially reflect how real estate transactions may proceed when compared to the more prevalent practice of using neighborhoods as shorthand. Similarly, we document how models perform in response to explicit racial identity to initially understand the phenomenon more broadly. More realistically, race is conveyed through proxies or retained user composite profiles in persistent memory (Tonneau et al. 2026), and further studies should seek to address this methodologically. We also tested with internet access disabled to evaluate the underlying foundation models. Finally, although justification terms were collected for topic modeling, analysis per prompt condition across cities, identities, and other variables made topic modeling, at this stage of the work, a more dedicated analytic project.

Conclusion

As large language models increasingly seek to become “everything apps”—the aperture for new forms of information aggregation (Google 2026), internal integration of app-based searches (OpenAI 2026a), and external transactions across domains (OpenAI 2026b)—their function becomes that of an intermediary. Language models in an intermediary role interpret the user’s identity and preferences, filter what information and options become salient, and inform user visibility and access to what information and choices are relevant. In the housing context, this performed role is not neutral. The data LLMs learn from reflect racialized geographies that fair housing law has spent decades working to dismantle, not the reparative trajectory that institutions aspire to and work toward.

Our findings highlight that steering is not an intrinsic property of a model, but an emergent systems-interaction behavior mediated by the interaction of user identity and stated preference, in relation to the spatial logic of a specific urban geography. The magnitude of potential steering behavior was varied across different cities, suggesting that LLM recommendation behaviors are highly contextualized to underlying geographic and socioeconomic characteristics. Similarly, these behaviors also are responsive to the identity of the user and their preferences, further complicating LLM evaluation in place-based sectors like housing. Con-

textual mitigation approaches behave differently for various demographic groups, reflecting that models do not identically reproduce demographic patterns; rather, they interpret them, filtering what a city offers through what they assume a given individual should want and where they belong. Without deliberate intervention and critical domain judgment exercised at the real estate practitioner level, LLMs risk being a new mechanism for the reproduction of inequitable patterns through atomized, personalized recommendations and framing. Future work should continue to examine how proxy data, persistent memory, and implicit racialized cues affect search and recommendation more broadly in LLMs, and what governance frameworks are sufficient for the information intermediary role that conversational AI is beginning to play in everyday life.

Acknowledgments

The authors thank Jameel Khan, Andrea Lau, and former colleague Laurie Benner for their computational analysis and visualizations, feedback, and engagement that helped this study.

References

- Abid, A.; Farooqi, M.; and Zou, J. 2021. Persistent Anti-Muslim Bias in Large Language Models. ArXiv:2101.05783 [cs.CL].
- Barocas, S.; and Selbst, A. D. 2016. Big Data’s Disparate Impact. *California Law Review*.
- 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. Virtual Event Canada: ACM. ISBN 978-1-4503-8309-7.
- Bertrand, M.; and Mullainathan, S. 2004. Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination. *American Economic Review*, 94(4): 991–1013.
- Black, E.; Koepke, J. L.; Kim, P.; Barocas, S.; and Hsu, M. 2023. Less Discriminatory Algorithms. *SSRN Electronic Journal*.
- Butler, D. M.; and Crabtree, C. 2021. Audit studies in political science. In *Advances in Experimental Political Science*, 42–55. Cambridge University Press.
- Chen, J.; de Jong, K.; Poole, A.; Burakowski, J.; Nosti, E. E.; Windt, J.; and Wang, C. 2026. Uncovering Political Bias in Large Language Models using Parliamentary Voting Records.
- Choi, A.; Herbert, K.; and Winslow, O. 2019. Long Island Divided. Technical report, Newsday.
- Colfax, R. 2024. Fair Lending Monitorship of Upstart Network’s Lending Model. Technical Report 4.
- Congress, U. S. 1968. 42 USC Ch. 45: FAIR HOUSING.
- Dudy, S.; Tholeti, T.; Ramachandranpillai, R.; Ali, M.; Li, T. J.-J.; and Baeza-Yates, R. 2025. Unequal Opportunities: Examining the Bias in Geographical Recommendations by

- Large Language Models. In *Proceedings of the 30th International Conference on Intelligent User Interfaces*, IUI '25, 1499–1516. New York, NY, USA: Association for Computing Machinery. ISBN 979-8-4007-1306-4.
- Google. 2026. A new era for AI Search.
- Graham, M. 2010. Neogeography and the Palimpsests of Place: Web 2.0 and the Construction of a Virtual Earth. *Tijdschrift voor Economische en Sociale Geografie*, 101(4): 422–436. eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-9663.2009.00563.x>.
- Graham, M.; and Dittus, M. 2022. *Geographies of Digital Exclusion: Data and Inequality*. Pluto Press, 1 edition. ISBN 978-1-78680-741-0 978-0-7453-4019-7.
- Graham, M.; and Zook, M. 2013. Augmented Realities and Uneven Geographies: Exploring the Geolinguistic Contours of the Web. *Environment and Planning A: Economy and Space*, 45(1): 77–99.
- Hangen, F.; and O'Brien, D. T. 2023. The Choice to Discriminate: How Source of Income Discrimination Constrains Opportunity for Housing Choice Voucher Holders. *Urban Affairs Review*, 59(5): 1601–1625.
- Hartmann, J.; Schwenzow, J.; and Witte, M. 2023. The political ideology of conversational AI: Converging evidence on ChatGPT's pro-environmental, left-libertarian orientation. *SSRN Electronic Journal*.
- Housing and Urban Development, D. o. 2013. Implementation of the Fair Housing Act's Discriminatory Effects Standard.
- Housing and Urban Development, D. o. 2026. HUD's Implementation of the Fair Housing Act's Disparate Impact Standard.
- Jang, K. M.; Chen, J.; Kang, Y.; Kim, J.; Lee, J.; Duarte, F.; and Ratti, C. 2024. Place identity: a generative AI's perspective. *Humanities and Social Sciences Communications*, 11(1): 1156.
- Kerche, F. W.; Zook, M.; and Graham, M. 2026. The silicon gaze: A typology of biases and inequality in LLMs through the lens of place. *Platforms & Society*, 3: 29768624251408919.
- Lawyers' Committee. 2025. Case Update: D.C. Court Allows ERC's Lawsuit Against Meta Over Algorithmic Bias to Proceed.
- Li, B.; Haider, S.; and Callison-Burch, C. 2024. This Land is Your, My Land: Evaluating Geopolitical Bias in Language Models through Territorial Disputes. In Duh, K.; Gomez, H.; and Bethard, S., eds., *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 3855–3871. Mexico City, Mexico: Association for Computational Linguistics.
- Liu, E. J.; So, W.; Hosoi, P.; and D'Ignazio, C. 2024. Racial Steering by Large Language Models: A Prospective Audit of GPT-4 on Housing Recommendations. In *Proceedings of the 4th ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, EAAMO '24, 1–13. New York, NY, USA: Association for Computing Machinery. ISBN 979-8-4007-1222-7.
- Manvi, R.; Khanna, S.; Burke, M.; Lobell, D.; and Ermon, S. 2024. Large Language Models are Geographically Biased. ArXiv:2402.02680 [cs.CL].
- Olin, A. 2020. Mapping low-income displacement and poverty concentration in Houston | Kinder Institute for Urban Research.
- OpenAI. 2026a. Introducing apps in ChatGPT and the new Apps SDK.
- OpenAI. 2026b. A new personal finance experience in ChatGPT.
- Peng, J.; Strijker, D.; and Wu, Q. 2020. Place Identity: How Far Have We Come in Exploring Its Meanings? *Frontiers in Psychology*, 11: 294.
- Proshansky, H. M. 1978. The City and Self-Identity. *Environment and Behavior*, 10(2): 147–169.
- Salinas, A.; Haim, A.; and Nyarko, J. 2024. What's in a Name? Auditing Large Language Models for Race and Gender Bias.
- Samad, H.; Lam, T.; Mügge-Durum, C.; and Akinwumi, M. 2026. The Geography of Algorithmic Judgment: LLM Intermediaries, Place Identity, and Racial Steering in Housing Search. ArXiv:2606.06694 [cs.LG] version: 2.
- Tonneau, M.; Seghal, N. K. R.; Malhotra, N.; Kazemi, S.; Orozco-Olvera, V.; Boudet, A. M. M.; Subramanian, L.; Fraiberger, S. P.; Guntuku, S. C.; and Hofmann, V. 2026. Different Demographic Cues Yield Inconsistent Conclusions About LLM Personalization and Bias. Version Number: 2.
- Turner, A.; Ross, S. L.; Galster, G. C.; Yinger, J.; Godfrey, E.; Bednarz, B. A.; Herbig, C.; Lee, S. J.; Rezaul Hussain, A.; Zhao, B.; and Institute, T. U. 2002. Discrimination in Metropolitan Housing Markets: National Results from Phase I HDS 2000.
- United States Supreme Court. 1971. *Griggs v. Duke Power Co.* | 401 U.S. 424 (1971) | Justia U.S. Supreme Court Center.
- United States Supreme Court. 1973. *McDonnell Douglas Corp. v. Green*, 411 U.S. 792 (1973).
- United States Supreme Court. 2015. *Texas Department of Housing and Community Affairs v. Inclusive Communities Project, Inc.*, 576 U.S. 519 (2015).
- Williams, D.; Kennedy, D. J.; Lillywhite, J.; Poscablo, C. S.; Clarke, K.; Majeed, S. S.; Hagler, R. T.; Baldon, J. L.; Song, H. C.; and Flagg, K. A. 2022. *UNITED STATES OF AMERICA v. META PLATFORMS, INC.*