Political Analysis (2025), **0**, 1–9
doi:10.1017/pan.2025.10022

PA

LETTER

Nationally Representative, Locally Misaligned: The Biases of Generative Artificial Intelligence in Neighborhood Perception

Paige Bollen[1], Joe Higton[2] and Melissa Sands[3]

[1]Department of Political Science, Ohio State University, Columbus, OH, USA; [2]Department of Politics, New York University, New York, NY, USA; [3]Department of Government, London School of Economics, London, UK

Corresponding author: Melissa Sands; Email: mlsands@gmail.com

(Received 6 March 2025; revised 16 July 2025; accepted 21 July 2025)

Abstract

Researchers across disciplines increasingly use Generative Artificial Intelligence (GenAI) to label text and images or as pseudo-respondents in surveys. But of which populations are GenAI models most representative? We use an image classification task—assessing crowd-sourced street view images of urban neighborhoods in an American city—to compare assessments generated by GenAI models with those from a nationally representative survey and a locally representative survey of city residents. While GenAI responses, on average, correlate strongly with the perceptions of a nationally representative survey sample, the models poorly approximate the perceptions of those actually living in the city. Examining perceptions of neighborhood safety, wealth, and disorder reveals a clear bias in GenAI toward national averages over local perspectives. GenAI is also better at recovering relative distributions of ratings, rather than mimicking absolute human assessments. Our results provide evidence that GenAI performs particularly poorly in reflecting the opinions of hard-to-reach populations. Tailoring prompts to encourage alignment with subgroup perceptions generally does not improve accuracy and can lead to greater divergence from actual subgroup views. These results underscore the limitations of using GenAI to study or inform decisions in local communities but also highlight its potential for approximating "average" responses to certain types of questions. Finally, our study emphasizes the importance of carefully considering the identity and representativeness of human raters or labelers—a principle that applies broadly, whether GenAI tools are used or not.

Keywords: Computational methods; Observational studies; Natural language processing

Edited by: Daniel J. Hopkins and Brandon M. Stewart

The rapid evolution of Generative Artificial Intelligence (GenAI) has provided social scientists with powerful new tools. Existing research has focused on text-based applications, such as classifying data (e.g., Heseltine and Clemm von Hohenberg 2024; Mellon *et al.* 2024; Ornstein, Blasingame, and Truscott 2025) and simulating survey responses based on demographic profiles (e.g., Argyle *et al.* 2023; Kim and Lee 2024; Kozlowski, Kwon, and Evans 2024). We instead consider GenAI's "vision" capabilities,—where Large Language Models (LLMs) can respond to images. Just as LLMs increasingly dominate text analysis tasks, Large Multimodal Models (LMMs) are increasingly used in place of traditional computer vision algorithms (e.g. Bontempi *et al.* 2025; Luckey *et al.* 2020; Melegrito *et al.* 2024; Tselentis, Papadimitriou, and van Gelder 2023; Tukur *et al.* 2025). We assess LMMs as a potential solution to the challenge of systematically measuring subjective context. It is well established that people respond to subjective perceptions of their environments (Herda 2010; Laméris, Hipp, and Tolsma 2018; Lippmann 1922;

Semyonov *et al.* 2004; Wong *et al.* 2025). Yet, capturing these subjective perceptions systematically and at scale remains a major challenge (Wong *et al.* 2012, 2020). We explore whether the latest developments in GenAI can help circumvent this by providing a scalable proxy for human subjective evaluations.

We turn to street view images, which are increasingly used to study context (Hwang and Naik 2023; Hwang *et al.* 2023; LeVan 2020; Sampson and Raudenbush 1999). Using a crowdsourced dataset of street view images from Detroit, Michigan, we assess how well three leading LMMs align with a national sample and local sample of Americans on perceptions of wealth, safety, and disorder. Our approach yields a distribution of assessments for each image from five distinct sources, enabling both correlation comparisons across images and evaluations of GenAI versus humans separately for each image.[1] This dual approach highlights both relative performance (how GenAI assessments correlate with human perspectives overall) and absolute performance (how closely GenAI matches human appraisals on a per-image basis).

Our findings have important implications for using GenAI in social science research, and for labeling or rating tasks more broadly. First, when assessing street view images, LMMs most accurately reflect broad, national patterns rather than local nuances. This distinction is important because local residents provide insight into community-specific concerns that national assessments overlook. For locally salient issues such as safety and neighborhood conditions, resident perspectives remain essential, and GenAI evaluations are a poor substitute. Second, prompting LMMs to adopt locally or demographically tailored personas does not improve performance and, in many cases, diminishes it. This has implications for the applicability of "synthetic sampling" across research contexts. Third, we document demographic and geographic biases in GenAI responses across several policy-relevant themes. Many policy and practitioner use cases of LMMs blur the line between "objective" and "subjective" evaluation tasks, as illustrated by a 2024 European Union report describing potential applications of AI to categorize images or activities as "suspicious" and automatically report potential threats, public disturbances, and safety hazards (Europol 2024, 22–23). Our results suggest that user discretion is warranted, especially where categorization tasks are more subjective. We also focus on gender, which shapes perceptions of public spaces, especially around safety and risk (Alsharawy *et al.* 2021; Ouali *et al.* 2020). Finally, as researchers and practitioners increasingly rely on machine learning (ML) methods to extract insights from large datasets, we offer systematic evidence that the identities and representativeness of the labelers who create training data matter.

1. Data and Methods

Our image data come from Mapillary, an open-source platform for street view images. We randomly sample 85 images from Detroit and pre-process each image as described in Section S9 of the Supplementary Material. For each image in our sample, we obtain a distribution of ratings across neighborhood attributes from a nationally representative human sample, a Detroit representative human sample, and five vision- and text-enabled GenAI models: OpenAI's GPT-4o and GPT-4.1, Google's Gemini 1.5 Pro and Gemini 2.5 Pro, and Meta's open-source Llama 4 (see Section S1 of the Supplementary Material). Here, we present results from the best performing model, GPT-4o; see Tables SI1 and SI5 in the Supplementary Material for other results. To account for their non-deterministic nature, we query these models 30 times for every question-image pair.

In addition to GenAI evaluations, we field two human surveys: a U.S. nationally representative survey via Prolific and a Detroit representative survey conducted as part of the Detroit Metro Area Community Study (DMACS) at the University of Michigan. DMACS recruits a representative sample of approximately 2,430 Detroit residents (see Section S11 of the Supplementary Material). Detroit's unique history—most notably the rise and fall of the auto industry in the 20th Century—sets it apart from the US overall. Roughly three-quarters of residents identify as Black, and socioeconomic indicators fall

[1] Replication materials can be accessed at https://doi.org/10.7910/DVN/7RN8QO (Bollen, Higton, and Sands 2025). Further code for the LMM queries can be accessed at https://github.com/joehigton/GenAILocalBias.

below national averages. These features make it an ideal setting to explore how GenAI models align with national versus local perspectives.

In the national sample, 800 Prolific respondents each evaluated three randomly selected images. In the DMACS sample, 2,430 Detroit residents each assessed one randomly selected image. For each image, respondents answered questions about the featured area's daytime and nighttime safety, wealth, and disorder. In total, we collect 61,200 ratings per LMM ($n = 30$ per image-question pair) and 3,230 human-generated ratings (on average $n = 28$ per image-question pair for each sample) across the 85 images, providing a distribution of responses for each image-question.

2. Results

What does it mean for the outputs of an LLM to be aligned with the views of a population? We adopt two main approaches to analyzing the relationship between GenAI and human responses. In the first, we measure how well GenAI correlates with each human sample at the image level across repeated ratings. In the second, we conduct pairwise comparisons between the average GenAI rating and the average human rating, demonstrating how often the two diverge significantly. Figure 1 presents results from these approaches, focusing on comparisons between human evaluations and those of GPT-4o. Sections S3 and S4 of the Supplementary Material show results from Gemini 2.5 and Llama 4, respectively.

The top panel of Figure 1 summarizes the overall relationship, across images, between each sample pair. With each point representing ratings of a single image, it shows the degree of linear correlation between human samples' average perceptions and GPT-4o's average assessments, separately for each neighborhood attribute. The relationship between GPT-4o's assessments and the U.S. sample's perceptions tends to be positive and roughly linear. However, there is a high degree of variation between human samples and across neighborhood attributes. First, GPT-4o consistently shows a weaker correlation with the local Detroit sample than the national sample for all neighborhood attributes. Second, GPT-4o tends to perform worse on questions about safety and disorder than those about wealth. For neighborhood wealth, GPT-4o approximates both the national and Detroit sample well ($r_{US} = 0.83, r_{Detroit} = 0.73$). Large disparities emerge when comparing performance on the other neighborhood attributes: for safety and disorder assessments, GPT-4o's ratings align well with the national sample ($r_{US} = [0.8, 0.83]$) but range from very weak ($r_{Detroit} = 0.16$ for disorder) to moderate ($r_{Detroit} = [0.58, 0.6]$ for safety) correlations with the Detroit sample. These patterns persist when we explicitly prompt GPT to rate as though "you live in Detroit" ($r_{Detroit} = 0.15$ for disorder; $r_{Detroit} = [0.59, 0.60]$ for safety). Gemini and Llama produce the same pattern, though the correlations are typically weaker (see Table SI1 in the Supplementary Material).

The bottom panel of Figure 1 shows our second approach. For each image, we conduct a t-test on the difference in means between GPT-4o's assessments and the human assessments of each neighborhood attribute to determine whether the two are statistically different (at $p < 0.05$). This approach allows us to identify the overall direction of the GenAI's statistically significant bias. The beige area of the bars represents the proportion of images for which GPT-4o and human assessments are not statistically different from each other, while the brown represents where GPT-4o significantly overestimates relative to humans, and the coral where it underestimates.

Even in cases where the distribution of GPT-4o ratings strongly correlates with that of human sample ratings (as measured by correlation coefficients), its ratings of individual images tend to stray from those of humans. That is, GenAI is better at recovering relative, rather than absolute ratings.

For neighborhood wealth, despite relatively strong linear correlations, GPT-4o significantly under-estimates both U.S. and Detroit samples' ratings in 67% and 38% of the images, respectively. Prompting GPT-4o to respond as a Detroit resident worsens the latter to 55%.

Safety assessments show no clear pattern. GPT-4o significantly under-estimates the U.S. sample's daytime safety ratings in 39% of images, while significantly over-estimating nighttime safety in 19%. GPT-4o under-estimates daytime safety ratings of Detroit residents in 11% (38% when prompted with

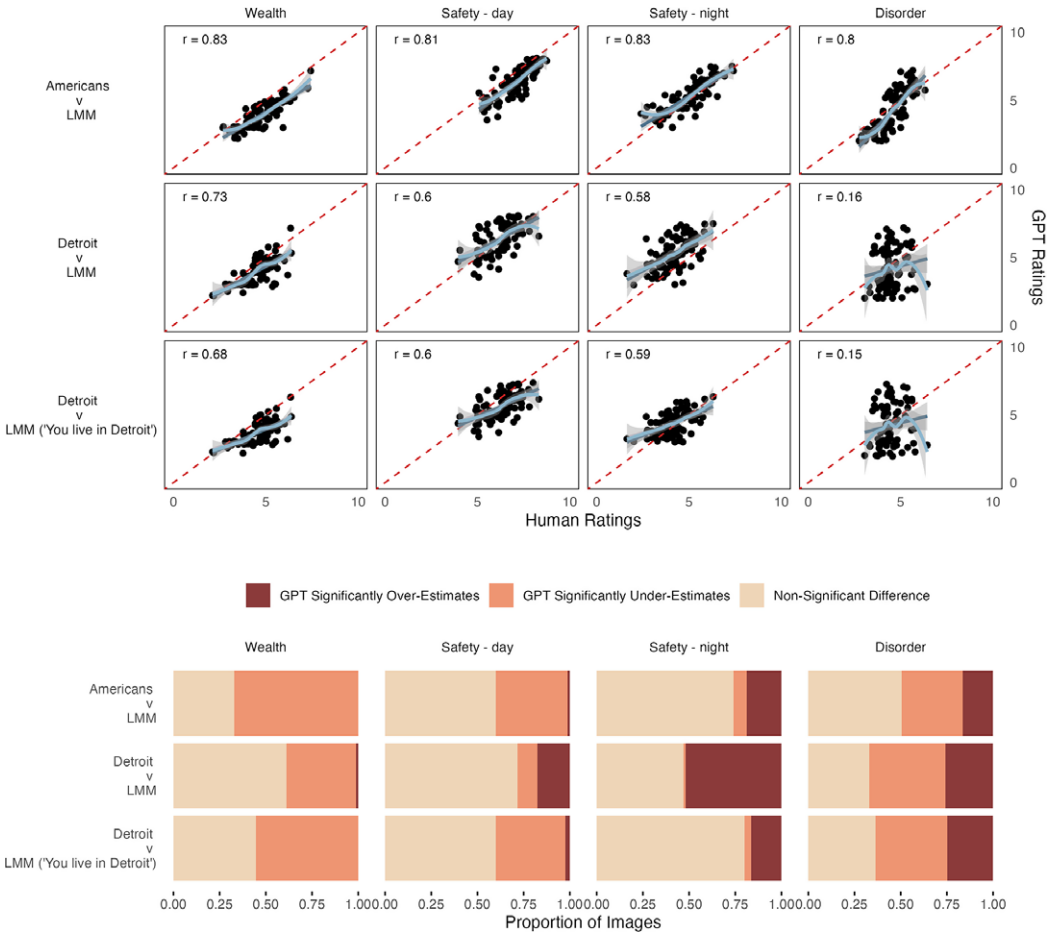

Figure 1. GPT's average evaluations of wealth, daytime safety, nighttime safety, and disorder compared to average evaluations of U.S. and Detroit samples.

Note: The top panel shows human samples' average perceptions plotted against GPT's average assessments. Each dot represents an image. The diagonal dashed line represents where the two are perfectly equivalent. Correlation coefficients and LOESS and linear regression lines with 95% confidence intervals are shown. The bottom panel displays the outcome of pairwise two-sample t-tests comparing the means of human- and GPT-derived ratings for each image.

"you live in Detroit") of images and over-estimates it in 18% (2% when prompted), but over-estimates nighttime safety in 52% (16% when prompted) and under-estimates it in 1% (4% when prompted).

Finally, the greatest disparities appear in disorder ratings, where GPT-4o significantly over- or under-estimates compared with the U.S. sample in 49% of images and with the Detroit sample in 67% (64% when prompted with "you live in Detroit").

Figures 2 and 3 visualize the extent to which LMM assessments align with those of humans across gender, which is known to be highly relevant to perceptions of neighborhoods (Alsharawy *et al.* 2021; Ouali *et al.* 2020). Because the (human) sample size is halved when looking within each gender, we focus on the correlation between the average ratings of GPT-4o and human respondents. Full t-test results are presented in Table SI1 in the Supplementary Material.

In the national sample, GPT-4o strongly aligns with both genders on safety. In the Detroit sample, GPT-4o consistently aligns more closely with women than with men across all neighborhood attributes, a trend especially pronounced for questions about safety. Gemini and Llama responses again follow a similar pattern, consistently correlating less well with human responses compared to GPT-4o while

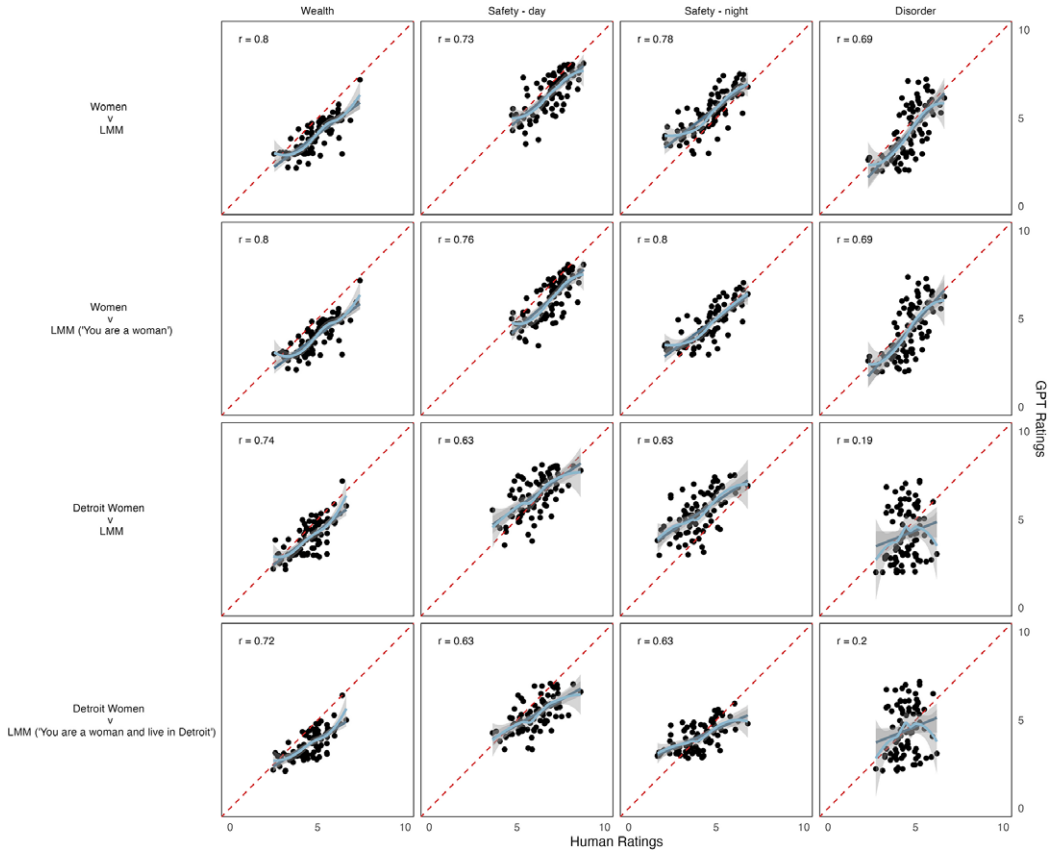

Figure 2. GPT's average evaluations of wealth, daytime safety, nighttime safety, and disorder compared to average evaluations of women in the U.S. and Detroit samples.
Note: The top panel shows human samples' average perceptions plotted against GPT's average assessments. Each dot represents an image. The diagonal dashed line represents where the two are perfectly equivalent. Correlation coefficients and LOESS and linear regression lines with 95% confidence intervals are shown.

displaying the same biases along gender and geographic lines (see Table SI1 and Figures SI2 and SI3 in the Supplementary Material). As before, prompt tailoring—asking the GenAI to adopt a particular identity when providing ratings—does little to improve correlations.

3. Discussion

We ask whose perspectives GenAI aligns with more closely: the generalized perspectives of the national sample or the localized, context-sensitive views of residents. By some metrics, GenAI is capable of broadly reflecting Americans' average perceptions of neighborhoods based on street view images. For some research or policy applications—those which require reasonable fidelity to a distribution of responses taken from a national sample—this may be sufficient. However, for other types of applications, its usefulness should be questioned.

First, where researchers seek subjective assessments of a small number of images, or where evaluations of individual images matter, GenAI can mislead. This includes use cases in the public sector, such as policing, as well as in the private sector, where in consumer research synthetic sampling is being used to evaluate visual stimuli such as images of product packaging (Sarstedt *et al.* 2024). We urge caution in use cases, where judgments based on individual images are important.

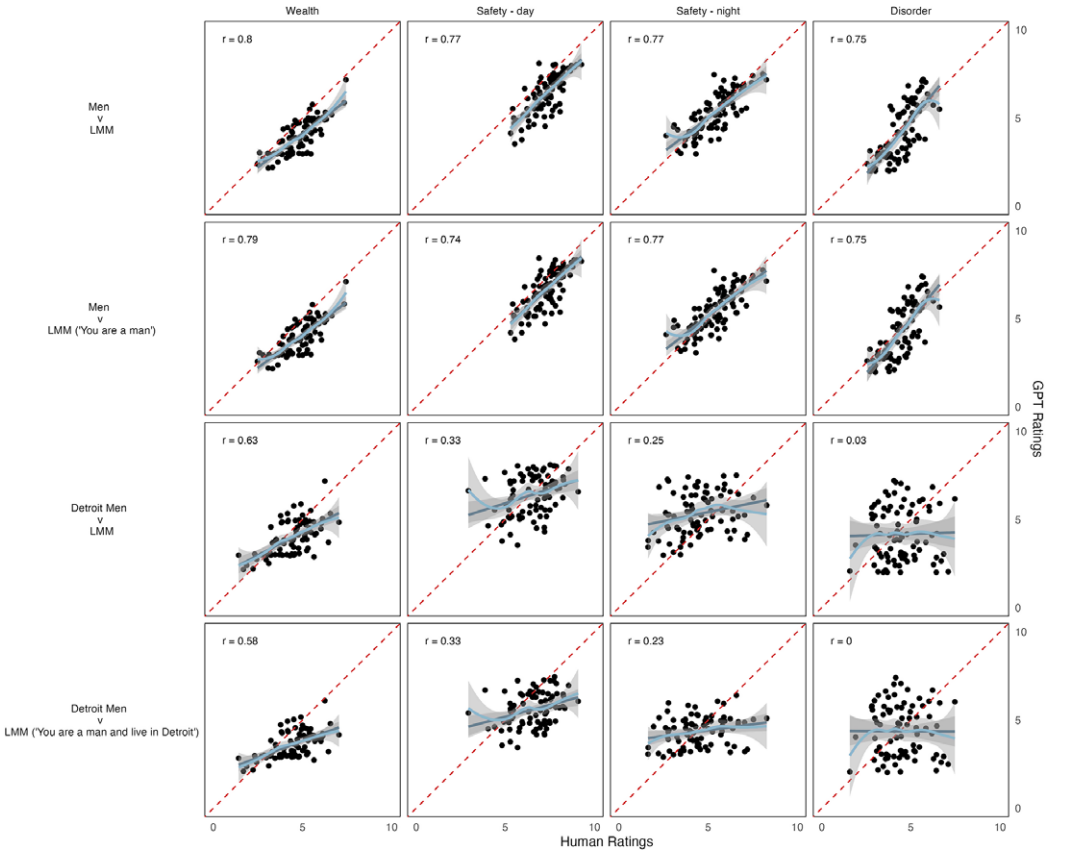

Figure 3. GPT's average evaluations of wealth, daytime safety, nighttime safety, and disorder compared to average evaluations of men in the U.S. and Detroit samples.
Note: The top panel shows human samples' average perceptions plotted against GPT's average assessments. Each dot represents an image. The diagonal dashed line represents where the two are perfectly equivalent. Correlation coefficients and LOESS and linear regression lines with 95% confidence intervals are shown.

Second, GenAI is unreliable where the goal is to approximate specific populations. Prompting LLMs to take the view of local residents does not address their biases and, in some cases, exacerbates them. We also find evidence that the more difficult-to-reach a population, the worse models perform at representing their views. In our tests, the largest disparities in the models' performance emerge between men and women residents of Detroit. This gap is most apparent when it comes to evaluations of safety, with GenAI aligning better with women than with men. Overall, GenAI is especially misaligned for men in Detroit, the majority of whom identify as Black or African American, a group with historically low research involvement (e.g., George, Duran, and Norris 2014). This finding is consistent with research that has demonstrated biases against African Americans in training data and in algorithms (e.g., Davidson, Bhattacharya, and Weber 2019; Mehrabi *et al.* 2021). However, Detroit's racial makeup does not explain our results. We show in Section S6 of the Supplementary Material that Black Detroiters' ratings are more similar to those of non-Black Detroiters than they are to those of Black U.S. residents (Table SI7 in the Supplementary Material). Moreover, GenAI aligns much more closely with assessments of non-Black U.S. residents than those of non-Black Detroiters (Table SI9 in the Supplementary Material).

One important limitation of our findings is that they merely reflect a "snapshot" of rapidly improving GenAI models. However, we have reason to believe that the ability of these models to reflect the voices

of hard-to-reach populations is unlikely to improve drastically, as survey data from these populations is scarce and typically incompatible with market incentives. We further show in Section S5 of the Supplementary Material that the latest iteration of the GPT model exhibits worse performance than its older counterpart. Given concerns over the reproducibility of GenAI models, we also detail in Section S10 of the Supplementary Material steps we took to address these issues as per Barrie, Palmer, and Spirling (2025).

Our focus on subjective assessments raises broader questions about how researchers should define "bias" and identify a relevant "ground truth" in image-based evaluations. In many social science applications of ML or GenAI, models are judged by how well they approximate an external standard. Yet in tasks involving subjective perceptions—such as detecting hateful speech or incivility online (Davidson *et al.* 2017; Southern and Harmer 2021), identifying trash on sidewalks (Hwang *et al.* 2023), or classifying aesthetics (Isola *et al.* 2013)—there may be no single, objective truth. What counts as "good quality" data in labeling? Perceived characteristics of environments shape outcomes independently of objective measures (Bowers *et al.* 2025; Gimpelson and Treisman 2018). In such cases, bias may mean deviation not from fixed reality, but from average human judgment—where "relevant" is defined by the use case. Our findings underscore the need to align model output with human perceptions when those perceptions drive behavior. As GenAI tools enter high-stakes contexts, researchers must ask not only what models get "right," but for whom—and by what standard.

Acknowledgments. The authors thank Ala' Alrababah, Daniel de Kadt, Elias Dinas, Liz Gerber, Ethan Porter, and audiences at Trinity College Dublin Department of Political Science Friday Seminar, the George Washington University American Politics Workshop, University of Barcelona IPERG Seminar Series, the Department of Politics and International Relations at Oxford University Comparative Politics Seminar, European University Institute and CIVICA Public Lecture Series Tours d'Europe, LSE Department of Methodology's Generative AI in Social Science Research Conference, the American Political Science Association 2024 Annual Meeting, and the Thought Summit on the Future of Survey Science at Cornell University for their invaluable feedback on early iterations of this project, and Muxuan Qu for research assistance.

Funding Statement. Data included in this report were collected and provided by the Detroit Metro Area Communities Study (DMACS). DMACS is a University of Michigan initiative that regularly surveys a broad, representative group of Detroit residents about their communities, including their experiences, perceptions, priorities, and aspirations. Support for DMACS comes from the University of Michigan Gerald R. Ford School of Public Policy, Institute for Social Research and Poverty Solutions. DMACS is also supported by the Knight Foundation, the Kresge Foundation and Ballmer Group. Learn more about DMACS at www.detroitsurvey.umich.edu and contact us at DMACS-info@umich.edu.

We received in-kind support from DMACS and a grant funded by the London School of Economics LSE Research Impact and Support Fund 2024.

Data Availability Statement. Replication materials can be accessed at https://doi.org/10.7910/DVN/7RN8QO (Bollen *et al.* 2025). Further code for the LMM queries can be accessed at https://github.com/joehigton/GenAILocalBias.

Author Contributions. P.B., J.H., and M.L.S. contributed equally to this work. Authors are listed in alphabetical order

Competing Interests. The authors declare none.

Ethical Standards. The research involving human subjects was reviewed and approved by the Harvard University Institutional Review Board (Protocol Number: IRB23-1746) and the London School of Economics Research Ethics Committee (Reference Number: 311794).

Supplementary Material. For supplementary material accompanying this paper, please visit https://doi.org/10.1017/pan.2025.10022.

References

Alsharawy, A., R. Spoon, A. Smith, and S. Ball. 2021. "Gender Differences in Fear and Risk Perception during the Covid-19 Pandemic." *Frontiers in Psychology* 12: 689467.

Argyle, L. P., E. C. Busby, N. Fulda, J. R. Gubler, C. Rytting, and D. Wingate. 2023. "Out of One, Many: Using Language Models to Simulate Human Samples." *Political Analysis* 31 (3): 337–351.

Barrie, C., A. Palmer, and A. Spirling 2025. "Replication for Language Models Problems, Principles, and Best Practice for Political Science." https://arthurspirling.org/documents/BarriePalmerSpirlingTrustMeBro.pdf

Bollen, P., J. Higton, and M. Sands 2025. "Replication Data for: Nationally Representative, Locally Misaligned: The Biases of Generative Artificial Intelligence in Neighborhood Perception." Harvard Dataverse. https://doi.org/10.7910/DVN/7RN8QO.

Bontempi, D., et al. 2025. "Faceage, a Deep Learning System to Estimate Biological Age from Face Photographs to Improve Prognostication: A Model Development and Validation Study." *The Lancet Digital Health* 7 (6): 100870

Bowers, J., C. Wong, D. Rubenson, M. Fredrickson, and A. Rundlett. 2025. "A Two Path Theory of Context Effects: Pseudoenvironments and Social Cohesion." Working Paper.

Davidson, T., D. Bhattacharya, and I. Weber. 2019. "Racial Bias in Hate Speech and Abusive Language Detection Datasets." In S. T. Roberts, J. Tetreault, V. Prabhakaran, & Z. Waseem (Eds.) *Proceedings of the Third Workshop on Abusive Language Online*, 25–35. Florence, Italy: Association for Computational Linguistics.

Davidson, T., D. Warmsley, M. Macy, and I. Weber. 2017. "Automated Hate Speech Detection and the Problem of Offensive Language." In D. Ruths (Ed.), *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 11, 512–515. Montreal, Canada: AAAI.

Europol. 2024. "AI and Policing: The Benefits and Challenges of Artificial Intelligence for Law Enforcement." Technical report, Europol Innovation Lab observatory report, Publications Office of the European Union, Luxembourg.

George, S., N. Duran, and K. Norris. 2014. "A Systematic Review of Barriers and Facilitators to Minority Research Participation among African Americans, Latinos, Asian Americans, and Pacific Islanders." *American Journal of Public Health* 104 (2): e16–e31.

Gimpelson, V., and D. Treisman. 2018. "Misperceiving Inequality." *Economics & Politics* 30 (1): 27–54.

Herda, D. 2010. "How Many Immigrants? Foreign-Born Population Innumeracy in Europe." *Public Opinion Quarterly* 74 (4): 674–695.

Heseltine, M., and B. Clemm von Hohenberg. 2024. "Large Language Models as a Substitute for Human Experts in Annotating Political Text." *Research & Politics* 11 (1): 20531680241236239.

Hwang, J., N. Dahir, M. Sarukkai, and G. Wright. 2023. "Curating Training Data for Reliable Large-Scale Visual Data Analysis: Lessons from Identifying Trash in Street View Imagery." *Sociological Methods & Research* 52 (3): 1155–1200.

Hwang, J., and N. Naik. 2023. "Systematic Social Observation at Scale: Using Crowdsourcing and Computer Vision to Measure Visible Neighborhood Conditions." *Sociological Methodology* 53 (2): 183–216.

Isola, P., J. Xiao, D. Parikh, A. Torralba, and A. Oliva. 2013. "What Makes a Photograph Memorable?" *IEEE Transactions on Pattern Analysis and Machine Intelligence* 36 (7): 1469–1482.

Kim, J. and B. Lee. 2024. "AI-Augmented Surveys: Leveraging Large Language Models and Surveys for Opinion Prediction." arXiv Preprint arXiv:2305.09620.

Kozlowski, A. C., H. Kwon, and J. A. Evans. 2024. "In Silico Sociology: Forecasting COVID-19 Polarization with Large Language Models." arXiv preprint arXiv:2407.11190.

Laméris, J., J. R. Hipp, and J. Tolsma. 2018. "Perceptions as the Crucial Link? The Mediating Role of Neighborhood Perceptions in the Relationship between the Neighborhood Context and Neighborhood Cohesion." *Social Science Research* 72: 53–68.

LeVan, C. 2020. "Neighborhoods that Matter: How Place and People Affect Political Participation." *American Politics Research* 48 (2): 286–294. Publisher: SAGE Publications Inc.

Lippmann, W. 1922. *Public Opinion*. New Brunswick, NJ: Transaction Publishers.

Luckey, D., H. Fritz, D. Legatiuk, K. Dragos, and K. Smarsly. 2020. "Artificial Intelligence Techniques for Smart City Applications." In E. Toledo Santos & S. Scheer (Eds.), *Proceedings of the 18th International Conference on Computing in Civil and Building Engineering: ICCCBE 2020*, 3–15. Springer.

Mehrabi, N., F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan. 2021. A Survey on Bias and Fairness in Machine Learning. *ACM Computing Surveys* 54 (6): 1–35.

Melegrito, M., et al. 2024. "Deep Learning Based Traffic Accident Detection in Smart Transportation: A Machine Vision-Based Approach." In H. Ogul & J. A. Morus (Eds.), *2024 4th International Conference on Applied Artificial Intelligence (ICAPAI)*, 1–6. IEEE.

Mellon, J., J. Bailey, R. Scott, J. Breckwoldt, M. Miori, and P. Schmedeman. 2024. "Do AIs Know What the most Important Issue Is? Using Language Models to Code Open-Text Social Survey Responses at Scale." *Research & Politics* 11 (1): 20531680241231468.

Ornstein, J. T., E. N. Blasingame, and J. S. Truscott 2025. "How to Train your Stochastic Parrot: Large Language Models for Political Texts." *Political Science Research and Methods* 13 (2): 264–281.

Ouali, L. A. B., D. J. Graham, A. Barron, and M. Trompet. 2020. "Gender Differences in the Perception of Safety in Public Transport." *Journal of the Royal Statistical Society Series A: Statistics in Society* 183 (3): 737–769.

Sampson, R. J., and S. W. Raudenbush. 1999. "Systematic Social Observation of Public Spaces: A New Look at Disorder in Urban Neighborhoods." *American Journal of Sociology* 105 (3): 603–651.

Sarstedt, M., S. J. Adler, L. Rau, and B. Schmitt. 2024. "Using Large Language Models to Generate Silicon Samples in Consumer and Marketing Research: Challenges, Opportunities, and Guidelines." *Psychology & Marketing* 41 (6): 1254–1270.

Semyonov, M., R. Raijman, A. Y. Tov, and P. Schmidt. 2004. "Population Size, Perceived Threat, and Exclusion: A Multiple-Indicators Analysis of Attitudes toward Foreigners in Germany." *Social Science Research* 33 (4): 681–701.

Southern, R., and E. Harmer. 2021. "Twitter, Incivility and "Everyday" Gendered Othering: An Analysis of Tweets Sent to UK Members of Parliament." *Social Science Computer Review* 39 (2): 259–275.
Tselentis, D. I., E. Papadimitriou, and P. van Gelder 2023. The Usefulness of Artificial Intelligence for Safety Assessment of Different Transport Modes. *Accident Analysis & Prevention* 186: 107034.
Tukur, H. N., O. Uwishema, H. Akbay, D. Sheikhah, and I. F. S. Correia. 2025. "Ai-Assisted Ophthalmic Imaging for Early Detection of Neurodegenerative Diseases." *International Journal of Emergency Medicine* 18 (1): 1–10.
Wong, C., J. Bowers, D. Rubenson, M. Fredrickson, and A. Rundlett. 2020. "Maps in People's Heads: Assessing a New Measure of Context." *Political Science Research and Methods* 8 (1): 160–168.
Wong, C., J. Bowers, D. Rubenson, M. Fredrickson, and A. Rundlett. 2025. "A Two Path Theory of Context Effects: Pseudoenvironments and Social Cohesion." Unpublished Manuscript.
Wong, C., J. Bowers, T. Williams, and K. D. Simmons. 2012. "Bringing the Person Back in: Boundaries, Perceptions, and the Measurement of Racial Context." *The Journal of Politics* 74 (4): 1153–1170.

Cite this article: Bollen, P., Higton, J. and Sands, M. (2025). Nationally Representative, Locally Misaligned: The Biases of Generative Artificial Intelligence in Neighborhood Perception. *Political Analysis*, 1–9. https://doi.org/10.1017/pan.2025.10022