What is Behind Homelessness Bias? Using LLMs and NLP to Mitigate Homelessness by Acting on Social Stigma

Jonathan A. Karr Jr. [1], Emory Smith[1], Matthew Hauenstein[1], Georgina Curto[2], Nitesh V. Chawla[1]
[1]University of Notre Dame, USA
[2]United Nations University Institute in Macau, Macau SAR, China
{jkarr, esmith36, mhauenst, nchawla}@nd.edu, curto@unu.edu

Abstract

Bias towards people experiencing homelessness (PEH) is prevalent in online spaces. This project will leverage natural language processing (NLP) and large language models (LLMs) to identify, classify, and measure bias using geolocalized data collected from X (formerly Twitter), Reddit, meeting minutes, and news media across the United States. While public opinion often refers to addictions, criminality, and high levels of welfare spending to justify bias against PEH, we will conduct a comparative study to determine whether racial fractionalization is associated with homelessness bias. The results of the study aim to provide a new path to alleviate homelessness by unveiling the intersectional bias that affects PEH and minority racial groups. During the course of the project, we will deliver a lexicon, compile an annotated database for homelessness and homelessness-racism intersectional (HRI) bias, evaluate LLMs as classifiers of homelessness and HRI bias, develop homelessness and HRI bias metrics, and audit existing LLMs on HRI. In collaboration with non-profits and the city council of South Bend, Indiana, USA, our ultimate goal is to contribute to homelessness alleviation by counteracting social stigma, restoring the dignity and well-being of the persons affected.

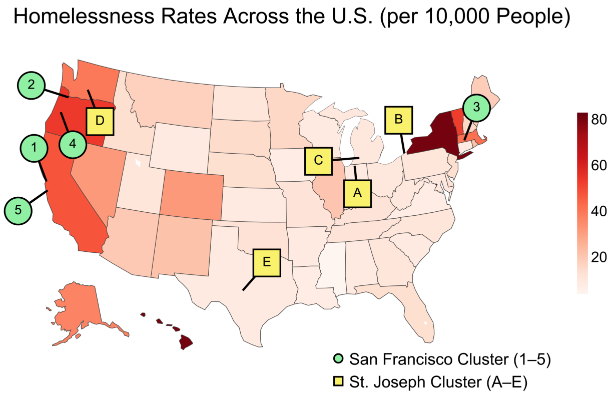
Homelessness Rates Across the U.S. (per 10,000 People)
1
2
3
4
5
A
B
C
D
E
80
60
40
20
● San Francisco Cluster (1–5)
□ St. Joseph Cluster (A–E)

Figure 1: This map shows two aspects: 1) the 2024 US Homelessness Rates per state from the Annual Homeless Assessment Report, and 2) the two sets of counties we will use for our comparative study. The counties marked with circular markers and numeric labels 1-5 are in the set of counties that includes San Francisco County. The counties marked with square markers and alphabetic labels A-E are the counties similar to St. Joseph County, Indiana, which includes the city of South Bend, Indiana. We will compare the 5 counties in each set to one another.

1 Introduction

The last time the United Nations conducted a survey of worldwide homelessness in 2005, they found that 100 million people were homeless [Kothari, 2005]. Nations with the largest homeless populations include Nigeria, Pakistan, and Afghanistan; each has over 4.5 million people experiencing homelessness (PEH) [World Population Review, 2024]. But the social challenge of homelessness is not limited to low-income countries. During one night in 2024, 771,480 people in the United States were recorded as experiencing homelessness, the highest number ever documented [de Sousa and Henry, 2024]. The homelessness crisis in the United States continues to escalate. From 2023 to 2024, Colorado, Alabama, Illinois, West Virginia, New York, and Massachusetts saw more than a 25% increase in the number of PEH [de Sousa and Henry, 2024]. This project focuses on the United States (Fig. 1), where previous studies have shown high indicators of bias against the poor [Curto *et al.*, 2024]. We aim to replicate and scale to an international level in future stages.

Various factors contribute to the homelessness crisis, and the stigmatization of PEH negatively impacts the mitigation of homelessness in many ways. Bias against the poor, especially the belief that the poor are undeserving, reduces peoples' support for crisis-mitigation policies [Applebaum, 2001]. Moreover, dedicated resources for homelessness mitigation consist mainly of providing shelter and covering basic material needs, while non-profits inform us that PEH consider social ostracization the worst consequence of being without a home [Caritas Spain, 2021]. Our project aims to contribute to the United Nations' first Sustainable Development Goal: Ending poverty in all its forms everywhere, by restoring the agency and capabilities of PEH [Sen, 2001].

Recognizing the relationship between stigmatization of PEH and efforts to reduce homelessness, our project will use natural language processing (NLP) and large language models (LLMs) to identify, classify, measure, and counteract stigma against PEH. We will present a comparative study between different counties in the US and examine the correlation between homelessness stigma and racial fractionalization. The counties in the scope of the project can be seen in Figure 1 and Table 1. Section 3.2 explains how the counties were chosen. In particular, we will inform, document and measure whether racism is associated with homelessness bias, as compared to the levels of criminality or high social spending, which have been found to be frequent arguments to consider PEH undeserving of help [Rex *et al.*, 2025].

This project is the result of a request from several nonprofit organizations that work daily on the urgent problems faced by PEH, including our partners in the community of South Bend, Indiana, USA. South Bend faces community conflicts due to the bias against PEH. In particular, the building of a new shelter has generated significant concern and controversy among local residents [Dellacca, 2024]. Our results will have a direct effect on this community and will inform other regions, both in the US and at a global level. Our ultimate goal is to open new paths to alleviate homelessness that take into account not only the need for practical and material solutions but also the dignity of the affected persons.

This paper is organized as follows: in Section 1, we present an overview of our project, the research questions, and expected outputs. In Section 2, we discuss the related work. In Section 3, we discuss the methods we will use to gather and annotate bias against PEH. We will use LLMs to annotate bias, and we will detect existing bias in LLMs. Section 4 is about the expected social impact. Section 5 describes the evaluation criteria. The project limitations are discussed in Section 6. We then include the implementation, scalability, and economic sustainability of our project, outlined in Section 7.We end with our ethical considerations.

The project aims to respond to two types of research questions (RQs):

1) Research questions aiming to contribute to pressing social challenges expressed by local stakeholders and domain experts on the topic of homelessness:

- RQ1 - How does homelessness bias vary across different US regions, and what role does racial fractionalization play in this variation?
- RQ2 - Can we apply our findings about homelessness-racism intersectional (HRI) bias to inform more effective policies that aim to alleviate homelessness?

2) RQs that contribute to the AI state of the art:

- RQ3 - How well can existing LLMs classify stigmatization of PEH and HRI bias, and how can their accuracy and performance be improved to meet classification standards?
- RQ4 - How biased are existing LLMs towards PEH (auditing)?
- RQ5 - How can we counteract homelessness and HRI bias using LLMs and NLP?

The project delivers the following research outputs (ROs):

- RO1 - Validated Lexicon on Homelessness and HRI: A lexicon on homelessness and the intersection between homelessness and racial minority groups in the USA, validated by domain experts and community stakeholders.
- RO2 - Multimodal Dataset on Homelessness and HRI: A dataset comprising geolocated data from X (formerly Twitter), Reddit, news media, and meeting minutes focusing on the discourse surrounding homelessness and considering the intersectionality with racial minority groups.
- RO3 - Annotated Corpus: A corpus for bias against PEH (including intersectionality with racism) and sentiment analysis, classified both manually by experts and automatically using LLMs.
- RO4 - Homelessness and HRI Bias Annotation Guidelines: Based on the exercise of manually annotating our dataset with the support of domain experts, we will deliver annotation guidelines to help the NLP community identify homelessness and HRI bias.
- RO5 - Thematic Analysis of Homelessness Discourse: Topic modeling applied to the dataset will reveal prevalent themes and patterns in online conversations about homelessness.
- RO6 - Novel HRI Bias Metric and Viability Analysis: Analysis of LLM accuracy on classifying homelessness and HRI bias. This will allow us to generate a homelessness bias index in future work to follow up on the public perception of homelessness throughout time and at an international level.
- RO7 - Geographic Visualization of Homelessness and HRI Bias: Map of homelessness bias across the United States.
- RO8 - Comparative Analysis Across Selected US Counties: Exploring the potential correlation between homelessness bias, racial fractionalization, and socioeconomic indicators, including crime rates and welfare spending.
- RO9 - LLM Audit of Bias For PEH and HRI: An evaluation and comparison of existing LLMs, auditing their biases, and identifying those exhibiting harmful language towards PEH.
- RO10 - Bias Mitigation Strategies for Online Textual Data: Tested strategies for reducing online bias, including text alteration techniques.
- RO11 - Policy Recommendations for Addressing Homelessness Stigma: Data-driven recommendations for policymakers to tackle homelessness by addressing social stigma, complementing traditional approaches focused on housing and basic needs [Marshall *et al.*, 2024].

2 Related Work

Our project will build on previous work regarding 1) using LLMs and NLP to identify patterns in bias against PEH and the poor, also known as aporophobia [Cortina, 2022], 2) the current limitations of LLMs when it comes to identifying bias, and 3) counteracting bias in LLMs and with LLMs.

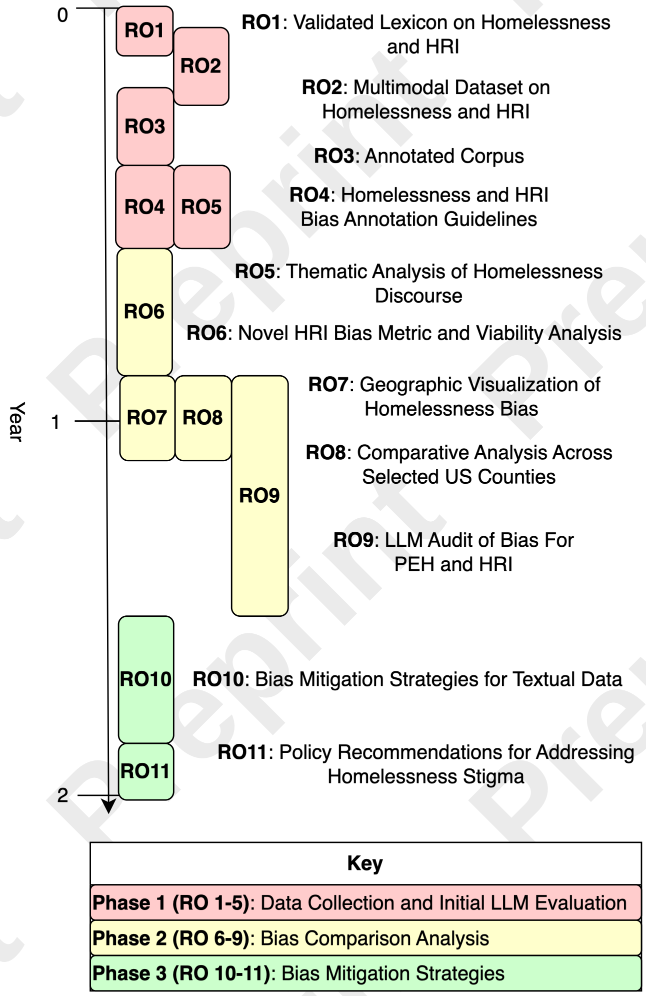
0
1
2
Year
RO1
RO2
RO3
RO4
RO5
RO6
RO7
RO8
RO9
RO10
RO11
RO1: Validated Lexicon on Homelessness and HRI
RO2: Multimodal Dataset on Homelessness and HRI
RO3: Annotated Corpus
RO4: Homelessness and HRI Bias Annotation Guidelines
RO5: Thematic Analysis of Homelessness Discourse
RO6: Novel HRI Bias Metric and Viability Analysis
RO7: Geographic Visualization of Homelessness Bias
RO8: Comparative Analysis Across Selected US Counties
RO9: LLM Audit of Bias For PEH and HRI
RO10: Bias Mitigation Strategies for Textual Data
RO11: Policy Recommendations for Addressing Homelessness Stigma
Key
Phase 1 (RO 1-5): Data Collection and Initial LLM Evaluation
Phase 2 (RO 6-9): Bias Comparison Analysis
Phase 3 (RO 10-11): Bias Mitigation Strategies

Figure 2: This is a two-year timeline which includes research outputs. As outlined in Section 7, the project is divided into three phases : (1) Data Collection and Initial LLM Evaluation; (2) Bias Comparison across US counties, homelessness, and HRI bias metrics; and (3) Bias Mitigation Strategies.

2.1 Using LLMs as Classifiers to Identify Online Patterns in Bias Against PEH and the Poor

Previous studies have used LLMs to classify and analyze online content considered biased against the poor [Kiritchenko *et al.*, 2023; Curto *et al.*, 2024; Rex *et al.*, 2025] and online content containing the term “homeless” [Ranjit *et al.*, 2024]. For example, an international comparative study was conducted on the criminalization of poverty in online public opinion [Curto *et al.*, 2024]. And, a taxonomy on bias against the poor, or aporophobia, has been proposed [Rex *et al.*, 2025]. Additionally, it has been shown that LLMs are able to detect changes in the attitudes towards PEH associated with socioeconomic factors [Ranjit *et al.*, 2024]. For example, according to tweets classified by LLMs, a larger population of unsheltered PEH correlates to more harmful generalizations about PEH [Ranjit *et al.*, 2024]. However, these previous studies have been limited by lexicons containing a single word, ‘homelessness’, or by collecting data from a single media source such as X (formerly Twitter).

2.2 Limitations of LLMs When It Comes to Bias against PEH

While previous studies have used LLMs to identify bias, this does not mean LLMs are fair or unbiased themselves. In fact, LLMs have been found to exhibit bias against socially stigmatized groups [Mei *et al.*, 2023; Bhutani *et al.*, 2024; Leidinger and Rogers, 2024], misportray and flatten the representations of demographic groups [Wang *et al.*, 2024], and are especially intersectionality biased across demographic dimensions [Lalor *et al.*, 2022]. When LLMs are used as standalone classifiers to identify biases, they have also been found to mislabel harmful language towards PEH [Ranjit *et al.*, 2024]. However, when LLMs are trained with expert annotations, they more accurately classify bias against PEH [Ranjit *et al.*, 2024]. We will build upon this work by auditing existing LLMs and using LLMs as classifiers after they have been trained on a dataset manually annotated by domain experts.

2.3 Counteracting Bias Against PEH in LLMs and with LLMs

To mitigate LLM toxicity and bias, previous research has explored a variety of techniques. They include data augmentation, data filtering and re-weighting, data generation, instruction tuning, projection-based mitigation, architecture modification, loss function modification, selective parameter updating, filtering model parameters, decoding, weight redistribution, and rewriting [Gallegos *et al.*, 2024]. Other approaches, such as safety system prompts, a form of instruction tuning, have been shown to reduce stereotyping and toxicity [Leidinger and Rogers, 2024]. Additional ways of mitigating bias include removing bias from word embeddings, augmenting training data, increasing the dropout rate in pretrained models, and fine-tuning pretrained models to encourage orthogonal representations [Lalor *et al.*, 2022]. Using identity-coded names instead of labels resulted in a more in-group portrayal from LLMs [Wang *et al.*, 2024]. After adding stylistic considerations to the model, adding information to LLM prompts was shown to destigmatize text about people struggling with substance use disorders [Bouzoubaa *et al.*, 2024]. Furthermore, LLMs as conversational moderators can provide specific and fair feedback on toxic behavior [Cho *et al.*, 2023].

Our project builds on this previous research by using techniques to identify and measure online bias with the support of LLMs and inform about bias against PEH. Moreover, we will offer new insights to better understand and mitigate this specific type of stigmatization and its intersectionality with racism. We aim to counteract the narrative of the “undeserving poor” [Applebaum, 2001] by providing empirical data.

3 Strategy and Method

3.1 Project Pipeline

This section outlines the pipeline for our project, explaining the steps we will take to answer RQ3, RQ4, and RQ5, and outputs RO1, RO2, RO3, RO4, RO5, RO6, RO9, and RO10.

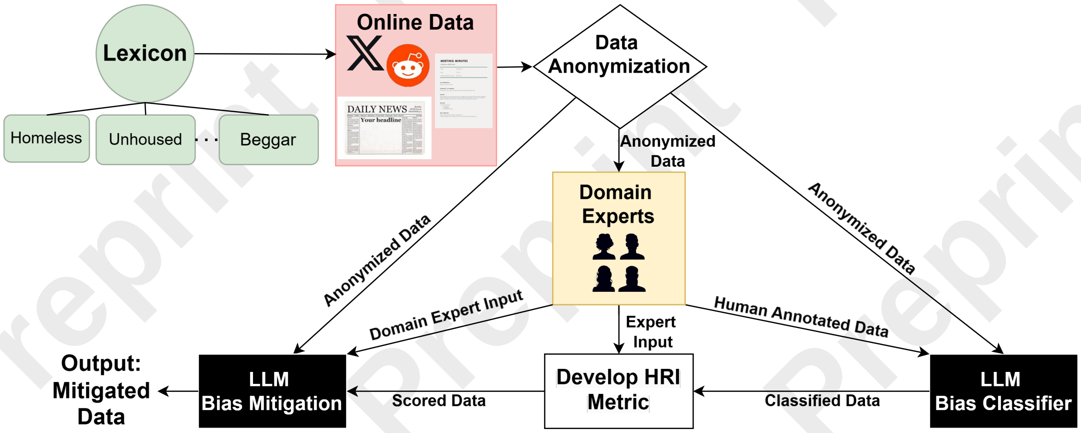

Figure 3: This shows the pipeline as described in Section 3.1. We start by using our lexicon to gather data from online sources. Then, we anonymize the data. Domain experts annotate a portion of the data to create a "Gold Standard". The "Gold Standard" is used to help a LLM classify the remaining data. Using the results, we will develop HRI metrics, and then come up with a bias mitigation strategy.

Obtaining the Data: Dataset Curation

We will generate a dataset of a minimum of 2,000 geolocalized manually annotated data items for homelessness bias, which will also document HRI bias, with the guidance of domain experts (RO2, RO3). The initial data will comprise approximately 500 items per source, drawn from X (formerly Twitter), Reddit, news media, and meeting minutes for counties across the United States. The selection of counties will be conducted based on socioeconomic demographic factors, including homelessness rates [United States Department of HUD, 2024] and racial fractionalization [U.S. Census Bureau, 2024; Alesina and Glaeser, 2013].

In order to generate the lexicon (RO1), we have compiled an initial list of keywords related to homelessness. This procedure involved utilizing LLMs and human experts. When using models such as GPT [Achiam *et al.*, 2023] or spaCy [Vasiliev, 2020], human experts evaluated what words were relevant. Words suggested by models with multiple use cases such as "shelter" were ignored since they could refer to "homelessness shelter" or other types of shelter. Our initial lexicon contains the words: 'homeless", "homelessness", "housing crisis", "affordable housing", "unhoused", "houseless", "housing insecurity", "beggar", "squatter", "panhandler", and "soup kitchen"..

Additionally, for a specialized analysis in the pilot project (South Bend, Indiana, USA), we will include a list of keywords tailored for the community. For example, keywords related to PEH in South Bend, include "Hotels4Now", "Center for the Homeless", "Hope Ministries South Bend", "St. Joseph County Homeless Coalition", "Real Services South Bend", "St. Vincent de Paul Society", and "Our Lady of the Road". Similar county-specific lexicons will be gathered for the various counties in our dataset. Once the data is gathered, we will anonymize the data to ensure privacy while preventing posts from being tracked back to users.

Identifying Homelessness and HRI Bias: Dataset Manual Annotation

We propose a two-level manual annotation for the data. The primary level classifies the text as "direct" or "reporting" bias [Rex *et al.*, 2025]. The second level annotation will be performed based on the topics found by the topic modeling on our initial collected data (RO5). We expect the topics will be similar to the topics found by previous studies, and we plan to use those previously found topics to validate our categories [Ranjit *et al.*, 2024].

Special attention needs to be paid to sentences that have multiple types of bias. For example, the sentence "Homeless people are always begging for money on the corner and scaring away my customers" has multiple types of bias, referring to PEH both as "being scary" and as "always begging".

The annotated dataset (RO3) will constitute our "Gold Standard" to determine the efficacy of LLMs in detecting homelessness bias. The manually annotated data will be balanced across the counties in scope. While annotating the data, domain experts will develop annotation guidelines to guide their own annotation and future annotations (RO4).

Sentiment Analysis

We also propose to conduct a sentiment analysis of the anonymized dataset to complement the geographic comparative study across the United States. This will allow us to document if, when referring to homelessness, the collected data items are classified as positive, neutral/objective, or negative. To accomplish this, we will use algorithms such as Naive Bayes [Parveen and Pandey, 2016] on our anonymized data set.

Towards a Homelessness Bias Index: LLMs as Classifiers

Annotating data manually is labor intensive, not sustainable, and not conducive to generating an ongoing homelessness and HRI bias index that is scalable at an international level. Once we have created an initial sample dataset of annotated data, we will evaluate the efficiency and efficacy of LLMs (including GPT, LLaMA, Claude, and Gemini) to detect both homelessness bias and HRI (RQ3).

Map Key	County, State (City Within County)	RFI*	Population	RPP†	RPA‡	Homelessness▽	GINI×
	Counties / Cities Comparable to San Francisco County (San Francisco, CA, USA)						
1	San Francisco County, California (San Francisco)	0.75	851,036	1032	131	98	0.52
2	Multnomah County, Oregon (Portland)	0.56	808,098	1198	237	91	0.47
3	Hampden County, Massachusetts (Springfield)	0.63	464,575	1534	182	65	0.47
4	Lane County, Oregon (Eugene)	0.40	382,218	1578	162	81	0.46
5	Santa Cruz County, California (Santa Cruz)	0.67	268,571	1091	96	69	0.48
	Counties / Cities Comparable to St. Joseph County (South Bend, IN, USA)						
A	St. Joseph County, Indiana (South Bend)	0.52	272,388	1378	97	8	0.47
B	Erie County, Pennsylvania (Erie)	0.34	270,495	1466	152	17	0.46
C	Kalamazoo County, Michigan (Kalamazoo)	0.43	261,426	1297	83	25	0.46
D	Yakima County, Washington	0.70	256,143	1532	129	0	0.42
E	Hays County, Texas	0.64	245,351	1253	83	6	0.45

*RFI: Racial Fractionalization Index
†RPP: Rate of People Below Poverty Line (per 10k)
‡RPA: Rate of People With Public Assistance (per 10k)
▽Homelessness: Homelessness Rate (per 10k)
×GINI: Income Inequality (GINI)

Table 1: Table of US counties with socioeconomic indicators for analyzing correlations with bias against PEH. While marginalization of PEH has often been justified using levels of criminality and high welfare spending, we are examining whether racial fractionalization influences homelessness bias to unveil a potential homelessness-racism intersectional bias. These results would offer a new path to alleviate homelessness by acting on public opinion and restore the dignity of the persons affected by homelessness, who are often ostracized.

We propose two ways to evaluate LLMs as classifiers:

1. The selection of anonymized labeled data will be classified (as biased or not biased) using zero-shot and few-shot approaches by LLMs. The outputs will be compared against the "Gold Standard" dataset and we will obtain F1 and cF1 scores [Yacouby and Axman, 2020].
2. We will also include a portion of the "Gold Standard" dataset as training data for the LLMs. We will follow the standard procedure for the train-test split based on the size of our data [Tan *et al.*, 2021] and calculate F1 and cF1 scores accordingly [Yacouby and Axman, 2020].

Using the results from manual annotation, sentiment analysis, and LLM classifications, we will develop metrics for HRI bias (RO6). The results of the LLMs' accuracy to detect homelessness and HRI bias will indicate the feasibility of generating an ongoing international homelessness and HRI index. The index will allow the tracking of homelessness and HRI bias over time, correlate it with social and political events, and inform homelessness-alleviation policy-making.

Auditing Existing LLMs for Homelessness and HRI Bias

In addition to evaluating how well existing LLMs can identify bias against PEH and HRI bias, we will also evaluate if biased text is produced by LLMs (RQ4, RO9). We will conduct this evaluation by prompting the LLMs with statements and require the LLMs to "fill in the blank" or autocomplete the statement. For example, we will ask the LLMs to complete the question, "Why do homeless people ...?" [Leidinger and Rogers, 2024]. We will also ask the LLMs to complete statements like, "The homeless person ..." [Gallegos *et al.*, 2024]. Based on the criteria developed while annotating the "Gold Standard" dataset, we will evaluate and compare the LLMs in scope.

Homelessness and HRI Bias Mitigation in LLMs

We will also utilize prompt engineering to reduce homelessness bias, based on the classifications outlined in the "Gold Standard" dataset (RQ5, RO10). Similar methods have been used when comparing how different models reduce other types of social bias [Raj *et al.*, 2024]. We foresee that these models will not be perfect for mitigating bias against PEH. Therefore, we will assess which models perform the best and integrate other LLM methods where appropriate.

3.2 Comparative Study on Homelessness Bias Among US Counties

The comparative study on homelessness bias among US counties will allow us to answer RQ1 and outputs RO7 and RO8. Because we are interested in testing the hypothesis that racial fractionalization increases bias [Alesina and Glaeser, 2013], we have chosen counties that have similarities in confounding factors like homelessness and poverty levels but differences in racial fractionalization. The preliminary list of sociodemographic confounding factors we will account for are: population, poverty, homelessness, income inequality, and public assistance. We have collected data about these confounding factors as well as racial fractionalization using sources such as the United States Census Bureau [U.S. Census Bureau, 2024] and the U.S. Department of Housing and Urban Development's (HUD) Point-in-Time Estimates (for homelessness) [United States Department of HUD, 2024]. Then, we used K-means clustering [Likas *et al.*, 2003] and qualitative analysis to choose counties that had similar rates of confounding factors with different rates of racial fractionalization. The homelessness rate numbers may not fully capture the true homelessness rate at the county level because HUD's data is reported by the census Continuum of Care (CoC) regions rather than by county. Although we apply a CoC to county mapping [Byrne *et al.*, 2013], this process can introduce effects on the data.

We have identified two sets of counties to analyze and compare, as seen in Table 1. The first set of counties, which includes San Francisco County, has larger populations, high homelessness rates, and contains cities whose homelessness crisis is often discussed in public discourse. We chose to analyze this set of counties because we want to address the severity of the homelessness crisis. The second set of counties, which includes St. Joseph County, has smaller populations and lower rates of homelessness. We have chosen this second set because our local partners have expertise in South Bend, Indiana, which is located in St. Joseph County, and will be able to provide nuanced input on the data.

Once we obtain the geolocalized online homelessness and HRI bias results per county, we will examine whether there is a correlation with racial fractionalization, when the confounding factors are similar (RQ1). We will utilize this information to make policy recommendations (RQ2, RO11). In line with Alessina's hypothesis [Alesina and Glaeser, 2013], this evidence aims to open new paths for homelessness alleviation by acting on the stigmatization of marginalized groups, changing the narrative that justifies the discrimination against PEH.

4 Expected Social Impact

The project aims to offer new insights to understand the phenomenon of bias against PEH. While traditional efforts to alleviate homelessness have focused on meeting the basic material needs of PEH, increasing studies document that the need of affiliation, having the social basis of self-respect and being treated as a dignified human being are at least equally important to effectively respond to this urgent social challenge [Narayan and Petesch, 2002; Marshall *et al.*, 2024]. In that respect, the project expects to evaluate generally accepted assumptions that justify bias against PEH, (including its correlation between high public spending or high criminality), versus potential alternative associated factors, such as racial fractionalization [Alesina and Glaeser, 2013], and therefore, racism. We expect, therefore, to act on the alleviation of homelessness by tackling the public narrative about the phenomenon, which is an obstacle to policies that mitigate poverty [Applebaum, 2001]. We will provide empirical data on the intersectionality of homelessness bias, racial discrimination, and the marginalization of minority groups in the US. Our study changes the focus of homelessness mitigation to the whole social fabric [Comim *et al.*, 2020].

Secondly, this project expects to contribute to solving urgent social challenges in a pilot project in the city of South Bend, Indiana, USA. South Bend faces difficulties in alleviating homelessness that are representative of cities across the nation. Namely, the building of a new shelter in South Bend has been met with concern by local residents [Dellacca, 2024]. We will collaborate with experts in South Bend throughout our project. When we have a strategy for counteracting bias online, we will pilot our strategy in online communities frequented by South Bend residents. We will also compare South Bend's county to counties across the nation to provide insight for South Bend's policy-makers. Within this scalable local pilot, our goal is to help the local community understand what is behind homelessness bias and how to implement more effective policies.

Moreover, by working with local non-profits, who are an integral part of our research team, our project aims to tackle one of the main discriminatory consequences of social stigmatization: PEH are not given a voice. This lack of voice is echoed in a study compiling the experiences of over 40,000 impoverished individuals in 50 countries, which found that "The mere fact of being poor is cause of being isolated, left out, looked down upon, alienated, pushed aside and ignored" [Narayan and Petesch, 2002].

Finally, our work will help counteract homelessness bias online, reducing toxicity in LLMs and social networks. We will provide a lexicon and a manually annotated dataset that will serve the research community to better identify, classify, and mitigate this type of bias online and its intersection with racism. Additionally, because social stigma can lead to violence [Allport, 1954], mitigating bias online can contribute to reducing violence against PEH.

5 Evaluation Criteria

The following criteria will be used at each project stage:

1. Lexicon (RO1). Once validated by the domain experts and non-profit organizations, we will evaluate the pertinence of each term by considering its usage frequency [Mikolov *et al.*, 2013]. Finally, we will ensure the strength of the terms' associations with homelessness using Pointwise Mutual Information [Kiritchenko *et al.*, 2023].

2. Topic modeling (RO5). We will compare our topics with the those found by previous studies [Ranjit *et al.*, 2024; Curto *et al.*, 2024]. We expect to find similar topics, such as "Not in my backyard" and undeservingness [Ranjit *et al.*, 2024] and criminalization and association with drugs [Curto *et al.*, 2024]. We will also ensure topic coherence and diversity. We will use normalized pointwise mutual information to measure topic coherence [Bouma, 2009]. For topic diversity, we will use the percentage of unique words in the top 25 words of all topics [Dieng *et al.*, 2020].

3. Manual annotations (RO3, RO4). To evaluate our expert annotations, we will use qualitative and quantitative measures. For qualitative evaluation, the first subset of the annotated dataset will be carefully discussed with domain experts belonging to specialized non-profit organizations and government collaborators. Moreover, annotation guidelines will be generated from the first focus group discussions. For quanti-

tative measures, we will measure inter-annotator agreement.

4. LLMs as classifiers (RQ3). We will evaluate how well LLMs can classify our dataset by using the following metrics: accuracy, precision, recall, and F1 score [Hu and Zhou, 2024]. We will perform cross-validation by dividing our dataset into multiple subsets. We will report the mean and variance across all of our split datasets. Additionally, domain experts will evaluate samples from the dataset and qualitatively describe the LLM's performance.

5. Metrics on homelessness bias (RO6). After we obtain homelessness and HRI bias metrics online, we will compare them with existing models that detect harmful language. We will first use existing bias metrics, including Regard [Sheng *et al.*, 2019], Holistic Evaluation of Language Models (HELM) [Liang *et al.*, 2022], and HateBERT [Caselli *et al.*, 2020] to score samples from our dataset. Then, we will score the same samples using our bias metric. Human evaluators will determine which metrics are most accurate.

7. LLM auditing (RO9, RQ4). When auditing existing LLMs to determine how biased they currently are, we will use our own bias-annotated dataset and homelessness-bias detection guidelines, as well as the existing toxicity models mentioned above. To evaluate our results, we will again have human participants score the text produced by the LLMs and compare their scores to our bias annotation guidelines. We will also perform cross-validation and report the mean and variance across our samples.

8. Mitigation strategy (RO10, RQ5). After developing a strategy to counteract bias based on our findings, we will evaluate how well our strategy works with our bias metric, as well as additional metrics such as the rate of online community rule violations (noncompliance), toxicity, and level of engagement [Srinivasan *et al.*, 2019].

6 Limitations

Computational approaches that aim to contribute to complex social challenges have numerous limitations, and our project is not an exception. First, our data is not a random representation of the population in the selected counties within the United States. In fact, 42 million people in the United States do not have access to broadband internet at home [International Telecommunication Union, 2024; John Busby and Cooper, 2021], and we are just including a selection of online resources in our study. Furthermore, we are only using textual English data, so this data is not reflective of people who communicate online in other languages or in other modalities like video and photos. Finally, we can only analyze what people say online, which may or may not accurately reflect what they are feeling or thinking.

7 Implementation, Scalability, and Economic Sustainability

The project will span two years and is divided into three main phases: (1) Data Collection and Initial LLM Evaluation; (2) Bias Comparison across US counties, homelessness and HRI bias metrics; and (3) Bias Mitigation Strategies. The timeline can be found in Figure 2.

This project was awarded the Strategic Framework Grant from the University of Notre Dame, which guarantees the implementation of the described research plan and expected outputs. The project's findings and resources, including the annotated dataset, lexicon, and mitigation strategies, will be made publicly available, promoting reproducibility and wider adoption. The methodology developed in this project aims to be applied to homelessness bias throughout the world, as well as other forms of social bias. By focusing on open source tools and creating publicly available data, the project ensures long-term accessibility. The scope of this project focuses on specific regions of the United States, but future work aims to build on it at the international level.

By creating a diverse team of computer scientists, social scientists, and local community partners, we have laid a strong foundation for interdisciplinary collaboration. By working with non-profit organizations and the city council of South Bend, Indiana, our team collaborates with local experts to ensure the project's relevance and impact.

Ethical Statement

Gathering data and analyzing it as described in the project requires careful analysis from an ethical point of view. Despite working with domain experts and carefully evaluating our approaches, we recognize that bias detection is often not a black or white decision, and the data annotations made by our team may be biased. We also note that the conclusions of the aggregate data in the study do not represent every individual person within the geographic scope of the project. Previous work has discussed the risks and ethical issues associated with identifying and analyzing negative bias in online spaces [Hovy and Spruit, 2016], especially with toxic language detection [Vidgen *et al.*, 2019]. We will follow the ethical recommendations for the NLP domain, ensuring a human is in the loop for annotation processes, creating models with built-in explainability [Cortiz and Zubiaga, 2020], and mitigating overgeneralization [Hovy and Spruit, 2016].

Privacy recommendations are also relevant in this project. In line with the European AI Act framework [European Union, 2024] and the U.S. National Institute of Standards and Technology (NIST) AI Risk Management Framework [NIST, 2024], we will anonymize and sanitize the data, careful not to release any identifying information.

Additionally, we will discuss each step of the project with domain experts and stakeholders (including non-profit organizations) and address their concerns.

Acknowledgments

This project is a collaborative effort involving the City of South Bend, the University of Notre Dame Center for Social Concerns, local non-profits, and Dr. Margaret Pfeil, cofounder of a local non-profit called Motels4Now that provides housing for PEH. Together, these stakeholders have identified the need for a pilot project in the city of South Bend, Indiana, USA. We also thank the University of Notre Dame for the Strategic Framework Grant that makes this work possible.

References

[Achiam *et al.*, 2023] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[Alesina and Glaeser, 2013] Alberto Alesina and Edward L Glaeser. *Fighting Poverty in the US and Europe*. Oxford University Press, Oxford, 2013.

[Allport, 1954] Gordon W Allport. *The nature of prejudice*. Basic Books, 1954.

[Applebaum, 2001] Lauren D Applebaum. The influence of perceived deservingness on policy decisions regarding aid to the poor. *Political psychology*, 22(3):419–442, 2001.

[Bhutani *et al.*, 2024] Mukul Bhutani, Kevin Robinson, Vinodkumar Prabhakaran, Shachi Dave, and Sunipa Dev. Seegull multilingual: a dataset of geo-culturally situated stereotypes. *arXiv preprint arXiv:2403.05696*, 2024.

[Bouma, 2009] Gerlof Bouma. Normalized (pointwise) mutual information in collocation extraction. *Proceedings of GSCL*, 30:31–40, 2009.

[Bouzoubaa *et al.*, 2024] Layla Bouzoubaa, Elham Aghakhani, and Rezvaneh Rezapour. Words matter: Reducing stigma in online conversations about substance use with large language models. *arXiv preprint arXiv:2408.07873*, 2024.

[Byrne *et al.*, 2013] Thomas Byrne, Ellen A Munley, Jamison D Fargo, Ann E Montgomery, and Dennis P Culhane. New perspectives on community-level determinants of homelessness. *Journal of Urban Affairs*, 35(5):607–625, 2013.

[Caritas Spain, 2021] Caritas Spain. Relatorías Especiales de Naciones Unidas sobre la Extrema Pobreza y los Derechos Humanos y sobre el Derecho a una Vivienda Adecuada. Technical report, 2021.

[Caselli *et al.*, 2020] Tommaso Caselli, Valerio Basile, Jelena Mitrović, and Michael Granitzer. Hatebert: Retraining bert for abusive language detection in english. *arXiv preprint arXiv:2010.12472*, 2020.

[Cho *et al.*, 2023] Hyundong Cho, Shuai Liu, Taiwei Shi, Darpan Jain, Basem Rizk, Yuyang Huang, Zixun Lu, Nuan Wen, Jonathan Gratch, Emilio Ferrara, et al. Can language model moderators improve the health of online discourse? *arXiv preprint arXiv:2311.10781*, 2023.

[Comim *et al.*, 2020] Flavio Comim, Miháli Tamás Borsi, and Octasiano Valerio Medoza. The Multi-dimensions of Aporophobia. 2020.

[Cortina, 2022] Adela Cortina. Aporophobia. Why we reject the poor instead of helping them. *Princeton University Press, Princeton*, 2022.

[Cortiz and Zubiaga, 2020] Diogo Cortiz and Arkaitz Zubiaga. Ethical and technical challenges of ai in tackling hate speech. *The International Review of Information Ethics*, 29, 2020.

[Curto *et al.*, 2024] Georgina Curto, Svetlana Kiritchenko, Kathleen C Fraser, and Isar Nejadgholi. The crime of being poor: Associations between crime and poverty on social media in eight countries. In *Proceedings of the Sixth Workshop on Natural Language Processing and Computational Social Science (NLP+ CSS 2024)*, pages 32–45, 2024.

[de Sousa and Henry, 2024] Tanya de Sousa and Meghan Henry. The 2024 annual homelessness assessment report (ahar) to congress. Technical report, The U.S. Department of HUD, 2024.

[Dellacca, 2024] Aynslee Dellacca. 'this is a human rights issue': South bend community discusses homeless shelter placement, 2024. Accessed: 2025-01-20.

[Dieng *et al.*, 2020] Adji B Dieng, Francisco JR Ruiz, and David M Blei. Topic modeling in embedding spaces. *TACL*, 8:439–453, 2020.

[European Union, 2024] European Union. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (Artificial Intelligence Act) (Text with EEA relevance), 2024.

[Gallegos *et al.*, 2024] Isabel O Gallegos, Ryan A Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K Ahmed. Bias and fairness in large language models: A survey. *Computational Linguistics*, pages 1–79, 2024.

[Hovy and Spruit, 2016] Dirk Hovy and Shannon L Spruit. The social impact of natural language processing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 591–598, 2016.

[Hu and Zhou, 2024] Taojun Hu and Xiao-Hua Zhou. Unveiling llm evaluation focused on metrics: Challenges and solutions. *arXiv preprint arXiv:2404.09135*, 2024.

[International Telecommunication Union, 2024] International Telecommunication Union. Internet use continues to grow, but universality remains elusive, especially in low-income regions, 2024. Accessed: 2025-01-31.

[John Busby and Cooper, 2021] Julia Tanberk John Busby and Tyler Cooper. Broadbandnow estimates availability for all 50 states; confirms that more than 42 million americans do not have access to broadband, 2021. Accessed: 2025-01-31.

[Kiritchenko *et al.*, 2023] Svetlana Kiritchenko, Georgina Curto Rex, Isar Nejadgholi, and Kathleen C Fraser. Aporophobia: An overlooked type of toxic language targeting the poor. In *The 7th Workshop on Online Abuse and Harms (WOAH)*, pages 113–125, 2023.

[Kothari, 2005] Miloon Kothari. Report of the special rapporteur on adequate housing as a component of the right to an adequate standard of living. Technical report, UN Economic and Social Council Commision on Human Rights, 2005.

[Lalor *et al.*, 2022] John P Lalor, Yi Yang, Kendall Smith, Nicole Forsgren, and Ahmed Abbasi. Benchmarking intersectional biases in nlp. In *Proceedings of the 2022 conference of the North American chapter of the association for computational linguistics: Human language technologies*, pages 3598–3609, 2022.

[Leidinger and Rogers, 2024] Alina Leidinger and Richard Rogers. How are llms mitigating stereotyping harms? learning from search engine studies. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, pages 839–854, 2024.

[Liang *et al.*, 2022] Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*, 2022.

[Likas *et al.*, 2003] Aristidis Likas, Nikos Vlassis, and Jakob J Verbeek. The global k-means clustering algorithm. *Pattern recognition*, 36(2):451–461, 2003.

[Marshall *et al.*, 2024] Carrie Anne Marshall, Rebecca Gewurtz, Caitlin Ross, Alyssa Becker, Abrial, Laurence Roy Cooke, Rosemary Lysaght Skye Barbic, and Bonnie Kirsh. Beyond Securing a tenancy: using the capabilities approach to identify the daily living needs of individuals during and following homelessness. *Journal of Social Distress and Homelessness*, VOL. 33, N, 2024.

[Mei *et al.*, 2023] Katelyn Mei, Sonia Fereidooni, and Aylin Caliskan. Bias against 93 stigmatized groups in masked language models and downstream sentiment classification tasks. In *Proc. of the 2023 ACM FAccT*, pages 1699–1710, 2023.

[Mikolov *et al.*, 2013] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26, 2013.

[Narayan and Petesch, 2002] Deepa Narayan and Patti Petesch, editors. *Voices of the Poor. From Many Lands*. Oxford University Press & The World Bank, Washington, DC, 2002.

[NIST, 2024] NIST. Artificial intelligence risk management framework: Generative artificial intelligence profile. Technical Report NIST AI 600-1, U.S. Department of Commerce, 2024. Accessed: 2025-02-10.

[Parveen and Pandey, 2016] Huma Parveen and Shikha Pandey. Sentiment analysis on twitter data-set using naive bayes algorithm. In *2016 2nd iCATccT*, pages 416–419. IEEE, 2016.

[Raj *et al.*, 2024] Chahat Raj, Anjishnu Mukherjee, Aylin Caliskan, Antonios Anastasopoulos, and Ziwei Zhu. Breaking bias, building bridges: Evaluation and mitigation of social biases in llms via contact hypothesis. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, pages 1180–1189, 2024.

[Ranjit *et al.*, 2024] Jaspreet Ranjit, Brihi Joshi, Rebecca Dorn, Laura Petry, Olga Koumoundouros, Jayne Bottarini, Peichen Liu, Eric Rice, and Swabha Swayamdipta. Oathframes: Characterizing online attitudes towards homelessness with llm assistants. *arXiv preprint arXiv:2406.14883*, 2024.

[Rex *et al.*, 2025] Georgina Curto Rex, Svetlana Kiritchenko, Muhammad Hammad Fahim Siddiqui, Isar Nejadgholi, and Kathleen C Fraser. Tackling poverty by acting on social bias against the poor: a taxonomy and dataset on aporophobia. *Forthcoming at the Annual Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics*, 2025.

[Sen, 2001] Amartya Sen. *Development as freedom*. Oxford University Press, 2001.

[Sheng *et al.*, 2019] Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. The woman worked as a babysitter: On biases in language generation. *arXiv preprint arXiv:1909.01326*, 2019.

[Srinivasan *et al.*, 2019] Kumar Bhargav Srinivasan, Cristian Danescu-Niculescu-Mizil, Lillian Lee, and Chenhao Tan. Content removal as a moderation strategy: Compliance and other outcomes in the changemyview community. *Proc. of the ACM on HCI*, 3(CSCW):1–21, 2019.

[Tan *et al.*, 2021] Jimin Tan, Jianan Yang, Sai Wu, Gang Chen, and Jake Zhao. A critical look at the current train/test split in machine learning. *arXiv preprint arXiv:2106.04525*, 2021.

[United States Department of HUD, 2024] United States Department of HUD. 2007-2024 point-in-time estimates by coc, 2024. Accessed: 2025-02-10.

[U.S. Census Bureau, 2024] U.S. Census Bureau. American community survey 5-year data (2009-2023), 2024. U.S. Department of Commerce, Economics and Statistics Administration.

[Vasiliev, 2020] Yuli Vasiliev. *Natural language processing with Python and spaCy: A practical introduction*. No Starch Press, 2020.

[Vidgen *et al.*, 2019] Bertie Vidgen, Helen Margetts, and Alex Harris. How much online abuse is there. *Alan Turing Institute*, 11, 2019.

[Wang *et al.*, 2024] Angelina Wang, Jamie Morgenstern, and John P Dickerson. Large language models should not replace human participants because they can misportray and flatten identity groups. *ArXiv preprint, abs/2402.01908*, 2024.

[World Population Review, 2024] World Population Review. Homelessness by country 2024, 2024. Accessed: 2025-01-29.

[Yacouby and Axman, 2020] Reda Yacouby and Dustin Axman. Probabilistic extension of precision, recall, and f1 score for more thorough evaluation of classification models. In *Proceedings of the first workshop on evaluation and comparison of NLP systems*, pages 79–91, 2020.