Combating Homelessness Stigma with LLMs: A New Multi-Modal Dataset for Bias Detection

Jonathan A. Karr Jr.[1], **Benjamin F. Herbst**[1], **Ting Hua**[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, bherbst, thua, mhauenst, nchawla}@nd.edu, curto@unu.edu

Abstract

Homelessness is a persistent social challenge, impacting millions worldwide. Over 770,000 people experienced homelessness in the U.S. in 2024. Social stigmatization is a significant barrier to alleviation, shifting public perception, and influencing policymaking. Given that online and city council discourse reflect and influence part of public opinion, it provides valuable insights to identify and track social biases. This research contributes to alleviating homelessness by acting on public opinion. It introduces novel methods, building on natural language processing (NLP) and large language models (LLMs), to identify and measure PEH social bias expressed in digital spaces. We present a new, manually-annotated multimodal dataset compiled from Reddit, X (formerly Twitter), news articles, and city council meeting minutes across 10 U.S. cities. This unique dataset provides evidence of the typologies of homelessness bias described in the literature. In order to scale up and automate the detection of homelessness bias online, we evaluate LLMs as classifiers. We applied both zero-shot and few-shot classification techniques to this data. We utilized local LLMs (Llama 3.2 3B Instruct, Qwen 2.5 7B Instruct, and Phi4 Instruct Mini) as well as closed-source API models (GPT-4.1, Gemini 2.5 Pro, and Grok-4). Our findings reveal that although there are significant inconsistencies in local LLM zero-shot classification, the in-context learning classification scores of local LLMs approach the classification scores of closed-source LLMs. Furthermore, LLMs outperform BERT when averaging across all categories. This work aims to raise awareness about the pervasive bias against PEH, develop new indicators to inform policy, and ultimately enhance the fairness and ethical application of Generative AI technologies.

Content Warning: This paper presents textual examples that may be offensive or upsetting.

Code — https://github.com/Homelessness-Project/Multimodal-PEH-Classification

Dataset — https://zenodo.org/records/16877412

Introduction

Homelessness is a persistent social challenge that affects millions of people worldwide. The Organization for Economic Cooperation and Development (OECD) reports that there are 2.2 million people experiencing homelessness (PEH) in the OECD and EU countries (OECD 2024). The United States is no exception: more than 770,000 people were recorded as experiencing homelessness in 2024, the highest number ever documented (de Sousa and Henry 2024). Specifically, the Point in Time count for PEH in San Francisco alone increased by 52% between 2005 and 2024 (City and County of San Francisco 2024). In this context, there is a growing call for a shift from traditional homelessness management (which focuses on providing material resources) to comprehensive support approaches that also address the stigmatization of PEH (Union 2024).

The marginalization suffered by PEH remains an understudied topic (Rex et al. 2025). Biases against PEH contribute to dehumanizing those affected, and make it harder for policymakers to approve and implement social measures that aim to mitigate homelessness (Curto et al. 2024; Rex et al. 2025). Further, the public perception of homelessness influences public voting in elections and therefore has an impact on policies aimed at addressing it (Clifford and Piston 2017).

Online and city council discourse offer valuable insights into public opinion and the prevalence of social biases (Chan et al. 2021; Mislove et al. 2011). Leveraging these digital and public records presents an affordable and relatively rapid method to derive preliminary indicators of social biases expressed through language. This study contributes to the nascent field of research on agentic large language models (LLMs) for social impact. We present novel methods, building on natural language processing (NLP) and LLMs, to identify and measure bias against PEH expressed in these digital spaces. Our work explores the effectiveness of LLMs as classifiers for online and offline data to generate and track new indices of homelessness bias across various U.S. cities. We investigate potential correlations between these indices and explore avenues for tackling homelessness by influencing public opinion. To this end, we present the following research questions (RQs):

- **RQ1:** How well can existing LLMs classify stigmatization of PEH, and how can their performance of this task be improved?
- **RQ2:** How does English online and offline textual bias (identified in social networks and council meeting minutes, respectively) correlate with levels of homelessness in US counties?

- **RQ3:** How does English online textual homelessness bias differ across media platforms?

To answer these RQs, we accomplish the following tasks.

1. We collect and publish a dataset of online and offline geolocalized data on homelessness discourse between 2015 and 2025 for 10 US cities from Reddit, X (formerly Twitter), news articles, and city council meeting minutes.
2. We anonymize the data using spaCy.
3. We create a multi-modal PEH bias classification frame which expands upon previous studies(Ranjit et al. 2024; Rex et al. 2025)).
4. We classify biases against PEH in the multi-modal data using Local LLMs (Llama 3.2 3B Instruct, Qwen 2.5 7B Instruct, and Phi4 Instruct Mini), closed-source API models (GPT 4.2, Gemini 2.5 Pro, and Grok 4), and BERT, and compare them against human annotators.
5. Finally, we compare the identified bias across different cities and data sources using the best classification model, GPT-4.1. and highlight the social impact that bias against PEH can potentially have on the actual levels of homelessness.

Our approach aims to foster greater public awareness, reduce the spread of harmful biases, inform policy decisions, and ultimately enhance the fairness and ethical application of generative AI technologies in addressing social issues. Moreover, this study uses social data and LLMs to identify and measure social bias, with the goal of alleviating homelessness by acting on shared beliefs. We acknowledge the inherent risks associated with using AI to identify biases, particularly the potential for misclassifications (false positives or negatives) that could mislead public understanding. Therefore, this project is guided by the principle of beneficence, prioritizing the maximization of societal benefits while actively minimizing potential harms (Beauchamp 2008). To mitigate these risks and ensure the reliability of our AI models, we have created a human-annotated 'gold standard' dataset against which all models are compared. This gold standard was developed in close collaboration with domain experts from non-profit organizations and the City of South Bend, whose invaluable insights guided the identification and categorization of biases against PEH. Our partnership with the City of South Bend ensures that our research is not only academically sound but also practically relevant and actionable for policymakers on the ground.

Related Work

Understanding and addressing societal biases, particularly those against vulnerable populations, including PEH, is crucial for informing effective policy and fostering social equity. However, traditional social science methods for gauging public perception are often limited in their ability to process the large quantities of pertinent data available for analysis. Our research overcomes this constraint by leveraging advancements in AI, specifically LLMs and NLP, as powerful tools to systematically identify, measure, and track societal biases expressed in vast amounts of diverse textual data generated by humans. Therefore, we examine how current work (1) Evaluates and Benchmarks LLMs as Classifiers, addresses (2) Societal Impact and Policy-Oriented Data Collection, and uses (3) AI for Detecting and Classifying Societal Bias.

Evaluating and Benchmarking LLMs as Classifiers

Prior work benchmarks LLM capabilities in various classification tasks, particularly low-resource or novel scenarios like zero-shot and few-shot learning (Matarazzo and Torlone 2025). Studies evaluate their accuracy, consistency, and ability to generalize to new data distributions. For instance, benchmarks like GLUE and BIG-Bench, while general-purpose, offer foundational insights into core linguistic capabilities relevant for classification tasks (Wang et al. 2018; Srivastava et al. 2023). More holistically, HELM evaluates models across multiple dimensions, including fairness and bias, moving beyond mere accuracy (Liang et al. 2022).

While we focus on LLMs to detect human-generated bias, it is crucial to acknowledge the "inherent biases" within LLMs themselves (e.g., representational biases, harmful content generation) as these can influence classification outcomes (Li et al. 2025). Techniques for auditing LLM outputs for fairness across demographic groups or identifying stereotypical associations within their internal representations (Bolukbasi et al. 2016; Nadeem, Bethke, and Reddy 2020; Blodgett et al. 2020) are relevant for ensuring the integrity of the classification results we obtain. Recent work continues to investigate how LLMs inherit and manifest social biases from their training data, and how these biases can impact downstream applications like bias detection (Hartvigsen et al. 2022; Chaudhary 2024).

Societal Impact and Policy-Oriented Data Collection

NLP tools are being used to parse political activities, analyze legislation, track public sentiment, and investigate policy effects, transforming how researchers and policymakers engage with textual data (Jin and Mihalcea 2022). LLMs are proving valuable for tasks like coding large datasets, reducing reliance on manual annotation, and extracting meaningful information for policymaking (Gilardi, Alizadeh, and Kubli 2023; Halterman and Keith 2024; Li et al. 2024).

Research has also been done in mitigating biases within AI systems themselves (Morales, Clarisó, and Cabot 2024). The responsible application of AI in this context, including human-centered design principles, is critical to ensure that tools serve to reduce, rather than exacerbate, social disparities (Lu et al. 2024; UNESCO 2021).

AI for Detecting and Classifying Societal Bias

Previous studies have evaluated the effectiveness of LLMs as classifiers for biases against the poor, often collectively referred to as aporophobia, in online discourse (Kiritchenko et al. 2023; Curto et al. 2024; Rex et al. 2025). For instance, international comparative studies have shed light on the criminalization of poverty in online public opinion (Curto et al. 2024), and comprehensive taxonomies for aporophobia have been proposed (Rex et al. 2025).

More specifically concerning PEH, research has demonstrated LLMs' capability to detect shifts in public attitudes

linked to socioeconomic factors (Ranjit et al. 2024). For example, analyses of tweets classified by LLMs have indicated a correlation between a larger unsheltered PEH population and an increase in harmful generalizations (Ranjit et al. 2024). These pioneering efforts highlight the immense potential of computational methods for analyzing public sentiment and identifying societal biases at scale.

The OATH framework (Ranjit et al. 2024) provides one of the most comprehensive pipelines for homelessness bias classification, categorizing biases into nine frames: 'money aid allocation', 'government critique', 'societal critique', 'solutions/interventions', 'personal interaction', 'media portrayal', 'not in my backyard', 'harmful generalization', and 'deserving/undeserving'. However, OATH's data collection was limited to a single online platform (X, formerly Twitter) and relied on a single keyword ('homeless'). Our research significantly advances this area by collecting a novel, multimodal dataset from diverse online sources (Reddit, X, news articles) and, critically, incorporating offline data from city council meeting minutes, which offers unique insights into policy-level discourse. Furthermore, we utilize a comprehensive PEH Lexicon containing the words 'homeless', 'homelessness', 'housing crisis', 'affordable housing', 'unhoused', 'houseless', 'housing insecurity', 'beggar', 'squatter', 'panhandler', and 'soup kitchen' (Karr et al. 2025) and expand upon OATH's classification categories to capture a broader and more nuanced spectrum of biases, as detailed in Section .

Methodology

As noted in Figure 1, we create a multimodal dataset from Reddit, X, news articles, and meeting minutes by using the PEH lexicon (Karr et al. 2025). Then we anonymize the data with spaCy (Honnibal et al. 2020) to remove personally identifiable information (PII). We identify if the data contains bias against PEH and classify the types of biases using our multimodal PEH bias classification criteria. We use human annotators, LLMs, and BERT as PEH bias classifiers. We utilized local LLMs (Llama 3.2 3B Instruct, Qwen 2.5 7B Instruct, and Phi4 Instruct Mini) as well as closed-source API models (GPT-4.1, Gemini 2.5 Pro, and Grok-4), zero-shot and instruct(few-shot), and evaluated their performance against human annotators and BERT.

Data Collection

To collect the data, we selected 10 different cities in the US. Five of them are considered small in size and have low levels of homelessness, similar to South Bend, Indiana. We also select five larger cities similar to San Francisco, CA. Our code outlines how we created a list of 20 k-Nearest-Neighbors (kNNs) for the city list. When selecting cities, we filtered out those that had fewer than 50 Reddit posts on PEH between January 1st, 2015, and January 1st, 2025. The set of counties in Table 1 (corresponding to the selected set of cities) has similar levels of population, homelessness rates, and GINI, yet differs in racial fragmentation. We can also compare the differences in the two groups of cities since the San Francisco group contains larger cities and has higher levels of homelessness.

Multimodal Data Related to PEH	
Small Cities - Similar to South Bend, IN	
City	**County**
South Bend	St. Joseph County, IN
Rockford	Winnebago County, IL
Kalamazoo	Kalamazoo County, MI
Scranton	Lackawanna County, PA
Fayetteville	Washington County, AR
Large Cities - Similar to San Francisco, CA	
City	**County**
San Francisco	San Francisco, CA
Portland	Multnomah County, OR
Buffalo	Erie County, NY
Baltimore	Baltimore County, MD
El Paso	El Paso County, TX

Table 1: Counties Included in Multimodal PEH Analysis

We create a dataset on PEH by using the PEH lexicon (Karr et al. 2025) described in the Related Work. The breakdown of over 40,000 entities from Reddit, X, news articles, and city council meeting minutes between January 1st, 2015, and January 1st, 2025, is presented in Table 2. To scrape Reddit, we looked at the subreddits for each city. Since less than 3% of X posts are geolocized, we scraped data that was either geolocized or included the city by name. For news articles, we used the LexisNexis API.

Finally, we gathered data from city council meeting minutes. The cities in scope post their information in different ways, and two of the cities do not provide publicly accessible data. Seven cities have video or audio recordings that were transcribed via LLMs, while San Francisco provides the raw text.

Data Anonymization

Prioritizing the anonymization of our data is essential for research and privacy protection. We leveraged the capabilities of the spaCy NLP library (Honnibal et al. 2020). This technique allowed us to automatically identify and mask PII within the text. The specific categories of entities targeted for anonymization included: person name, geographic locations, organizations, and other identifying information such as street addresses, phone numbers, and emails. We also leveraged the Python module pydeidentify (Kogan 2023), which is based on spaCy, in case we missed any other information to be anonymized.

The result of this multifaceted anonymization strategy is a dataset that respects user privacy while retaining the essential content for bias analysis and the development of mitigation techniques.

Multimodal PEH Bias Classification Categories

We create categories for a multimodal PEH bias classification that has 16 categories and expands upon the nine OATH-Frames (Ranjit et al. 2024), noted in the Related Work. The OATH frames include different types of biases in discussion. However, the categories are limiting, since the

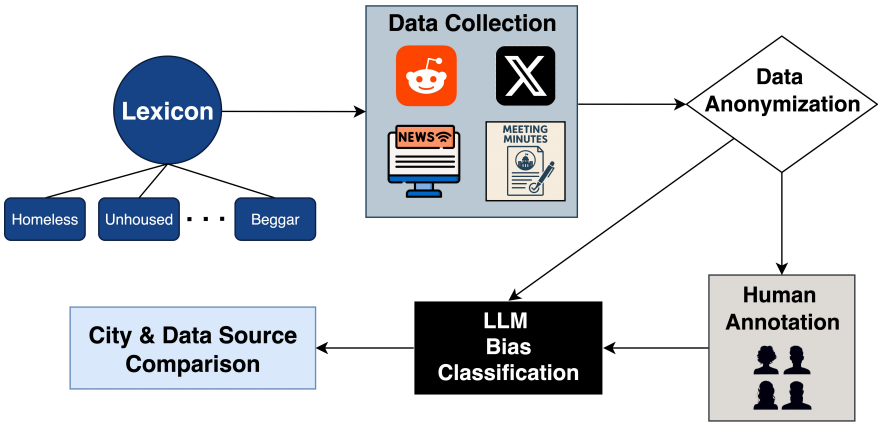

Figure 1: We collect Reddit, X, news articles, and city council offline meeting minutes data on homelessness discourse using the PEH lexicon (Karr et al. 2025). We then anonymize the data and have both LLMs and domain experts classify the data into different bias categories to determine the reliability of LLMs as classifiers.

City	Reddit		News		X (Twitter)			City Council	
	Posts	Comments	Articles	Paragraphs	Posts	Geolocated	Non-Reposts	Meetings	Comments
South Bend	62	**196**	36	**49**	96	6	**65**	86	**330**
Rockford	43	**188**	6	**9**	98	0	**43**	344	**243**
Kalamazoo	209	**1846**	8	**11**	99	1	**40**	N/A	N/A
Scranton	13	**79**	108	**159**	92	2	**56**	431	**514**
Fayetteville	34	**102**	28	**29**	97	3	**81**	233	**1043**
San Francisco	714	**14777**	1181	**1537**	9168	23	**2330**	25	**14**
Portland	751	**15301**	322	**397**	5574	39	**1215**	372	**6618**
Buffalo	151	**589**	176	**196**	685	1	**115**	211	**135**
Baltimore	246	**1215**	142	**156**	464	7	**244**	N/A	N/A
El Paso	40	**154**	28	**31**	99	1	**53**	74	**284**
Grand Total	2263	**34447**	2035	**2577**	16472	83	**4242**	3552	**9181**

Table 2: Summary of Data Across Different Cities

frames were designed for Twitter (now X) and do not include claims or questions as independent categories. Since questions are common on Reddit, we added the categories 'ask a genuine question' and 'ask a rhetorical question'. Additionally, we created the categories, 'provide a fact or claim' and 'provide an observation', based on the data we encountered through the annotation exercise. We also added the categories 'express their opinion' and 'express others' opinions', which indicate if the authors are giving a personal view or expressing the views of others. Finally, we include the category 'racist' that categorizes whether a post expresses racism or not.

Manually Annotated Baseline

Three human annotators labelled the dataset, using the defined multimodal PEH bias classification categories. We created a manually annotated baseline (Cardoso et al. 2014) utilizing stratified sampling (Liberty, Lang, and Shmakov 2016). To accomplish this, we annotated 50 comments per city (10 cities) for each of the four data sources. Since not all the cities had 50 entities for each source, a total of 1702 entities form our gold standard. When annotating, we worked in close collaboration with domain experts in the City of South Bend. This led us to have a high agreement rate among the human annotators, averaging 78.38% per category. However, it is not perfect since inevitably different people have different opinions about biases, based on their own personal experiences and backgrounds. Therefore, we construct the gold standard by utilizing soft labeling (Fornaciari et al. 2021), which takes an average of annotators' responses, and if two or more annotators agree, it is classified accordingly.

Results

Model Selection & Experimental Setup

To test and improve upon the current state of PEH bias classification, we benchmark a diverse set of models, encompassing both established deep learning architectures and state-of-the-art large language models (LLMs), against our human-annotated gold standard dataset. Our selection process was driven by the need to assess performance across

Data Source	GPT-4		LLaMA		Qwen		Phi-4		Grok		Gemini		BERT	
	Zero	Few	Zero	Few	Zero	Few	Zero	Few	Zero	Few	Zero	Few	Finetuned	–
Reddit (Macro)	75.00	**76.95**	64.92	59.94	66.09	70.58	60.62	63.35	60.05	61.98	60.67	63.47	37.43	
Reddit (Micro)	80.62	**82.93**	80.69	69.16	73.91	79.95	81.35	79.03	77.18	77.14	69.42	72.28	59.83	
X (Twitter) (Macro)	65.00	65.96	64.99	59.59	60.20	**70.98**	55.98	66.73	63.67	65.02	68.34	68.21	16.31	
X (Twitter) (Micro)	77.15	78.55	83.46	69.75	71.01	79.78	82.44	82.15	**83.69**	81.84	79.63	79.55	58.90	
News (Macro)	67.84	70.56	64.17	56.11	54.91	**73.02**	59.81	71.39	66.96	68.75	69.55	72.21	17.51	
News (Micro)	81.04	83.02	85.45	73.61	63.38	84.62	86.88	**87.06**	85.75	85.96	81.79	84.29	65.56	
Meeting Minutes (Macro)	66.59	70.50	65.67	61.49	64.31	**74.96**	60.31	63.97	66.56	70.70	70.87	73.10	21.45	
Meeting Minutes (Micro)	78.42	81.06	**84.89**	74.69	73.61	83.84	84.51	80.53	84.39	84.32	80.63	82.43	75.75	
Weighted Avg (Macro)	73.73	**75.78**	64.96	59.95	65.43	70.95	60.33	63.73	60.83	62.88	61.96	64.56	34.79	
Weighted Avg (Micro)	80.29	**82.56**	81.22	69.66	73.53	80.30	81.73	79.46	78.19	78.06	70.99	73.60	60.98	

Table 3: Macro and Micro F1 Scores for All Models by Data Source

different model sizes, architectures, and access modalities (local vs. API-based), as well as to investigate the impact of various prompting strategies.

We test six LLMs and BERT against our gold standard. We choose three local LLMs to determine how classification performs in low-resource environments: Llama 3.2 3B Instruct, Qwen 2.5 7B Instruct, and Phi4 Instruct Mini. We also test 3 closed-source LLMs: GPT-4.1, Gemini 2.5 Pro, and Grok-4. For all six LLMs, we used two types of prompts:

Zero-Shot Learning: This setup evaluates a model's inherent understanding and ability to classify unseen examples without any explicit task-specific examples in the prompt.

In-Context Learning (Few-shot): This strategy involves providing a small number of example input-output pairs directly within the prompt to guide the model's understanding of the task. For our experiments, we used five diverse examples from the human-annotated gold standard dataset for in-context learning. This process is standard in literature (Wang et al. 2020). These examples were chosen to represent a variety of sources and classification categories to maximize their utility.

Here are two examples of in-context prompts that the human annotators agreed on:

Example X: "Did your Black flunky mayor get the[image][ORGANIZATION]'s memo 2 stick it 2 Rump instead of serving you by refusing 2 deport migrants + give them Black taxpayers'[image] 4 shelter+food while Black citizens go homeless? [ORGANIZATION] mayors did. Charity starts at [image].[URL]"
Few-shot classification: 'ask a rhetorical question', 'provide a fact or claim', 'express their opinion', 'money aid allocation', 'harmful generalization', 'deserving/undeserving', and 'racist'.

Example Meeting Minutes: "but they stuck with us, they got all the permissions they needed, and we would not have made the functional end of veteran homelessness in [ORGANIZATION] without them, so thank you. PERSON0? Well, thank you for this honor."
Few-shot classification:'provide a fact or claim', 'express their opinion', and 'solutions/interventions'.

To assess model performance, we use macro-F1 score (Opitz and Burst 2019), yet also report the micro-F1 score. This metric is critical for multi-label classification tasks with potential class imbalance, as it calculates the F1 score for each individual class and then averages them, thus equally weighting the performance on both prevalent and rare categories. In our dataset, class imbalance is prevalent. For example, over 70% of the results in the gold standard are classified as 'provide a fact or claim', yet less than 1% are classified as racist.

We compare all models' overall F1 score on the gold standard for both zero-shot and few-shot. We then choose the best-performing model to test on the entire dataset. When choosing the best model, we pick the best macro-F1 score for the weighted average, where the weighted average is with respect to the number of results in the complete dataset.

Results and Analysis

In Table 3, we see that GPT has the best weighted average, so we chose it to classify our complete dataset. When analyzing the results, it is important to note that some models outperform GPT specifically for the X, news, and meeting minute tasks. Additionally, the local LLMs perform comparably to the closed-source models at these tasks. We also observe that BERT struggles with multi-classification, unlike LLMs. However, BERT micro-F1 score is significantly better than its macro-F1 score. This is because BERT finetunes to specific categories.

When examining Table 4, we find that few-shot learning helps improve categories that are underrepresented. However, zero-shot performance is better for categories where the LLM already performs well.

The correlation matrix in Figure 2 reveals that there are significant negative correlations between 'solutions/interventions' vs. 'societal critique', 'solutions/interventions' vs. 'harmful generalization', and 'solutions/interventions' vs. 'personal interaction'. There are also significant positive correlations between 'societal critique' vs 'deserving/undeserving', 'express their opinion' vs 'deserving/undeserving', and 'ask a genuine question' vs 'deserving/undeserving'. These findings can potentially help policymakers address homelessness alleviation by acting on public opinion. Our findings illustrate how, when there are harmful generalizations, there is less acceptance of solutions and interventions. Moreover, the positive correlation between "deserving / un-

Category	Reddit		News		Meeting Minutes		X (Twitter)	
	Zero	Few	Zero	Few	Zero	Few	Zero	Few
Ask Genuine Question	**78.95**	78.46	**12.90**	9.52	15.38	**18.18**	**13.16**	12.99
Ask Rhetorical Question	**73.22**	63.51	17.20	**22.50**	0.00	**22.22**	**28.57**	0.00
Deserving/Undeserving	6.67	**10.13**	4.55	**10.13**	0.00	**8.70**	0.00	**3.57**
Express Others Opinions	**39.34**	34.85	8.00	**14.71**	**19.23**	15.52	0.00	**2.74**
Express Opinion	**91.61**	91.06	**64.57**	64.17	25.64	**25.91**	63.98	**65.73**
Government Critique	**57.00**	56.22	**31.02**	28.40	15.00	**26.53**	16.39	**27.20**
Harmful Generalization	45.07	**48.70**	**25.33**	23.36	0.00	0.00	**21.62**	21.43
Media Portrayal	**9.30**	7.14	2.15	**4.76**	0.00	0.00	0.00	0.00
Money Aid Allocation	59.76	**60.44**	**29.85**	19.44	**35.64**	35.56	**35.29**	29.14
Not in My Backyard	50.53	**58.97**	**13.79**	0.00	0.00	0.00	**8.70**	0.00
Personal Interaction	54.84	**58.23**	**11.54**	8.00	0.00	0.00	**11.27**	9.76
Provide Fact/Claim	**80.50**	79.81	**83.18**	78.65	**93.69**	92.90	82.08	**89.06**
Provide Observation	53.81	**62.76**	6.40	**9.09**	3.39	**4.23**	**6.19**	4.04
Racist	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Societal Critique	**39.80**	38.39	19.79	**21.35**	13.33	**14.16**	**8.60**	6.90
Solutions/Interventions	68.75	**71.78**	43.15	**47.31**	52.52	**58.31**	51.43	**55.98**

Table 4: Category-wise F1 Scores for GPT4 Model

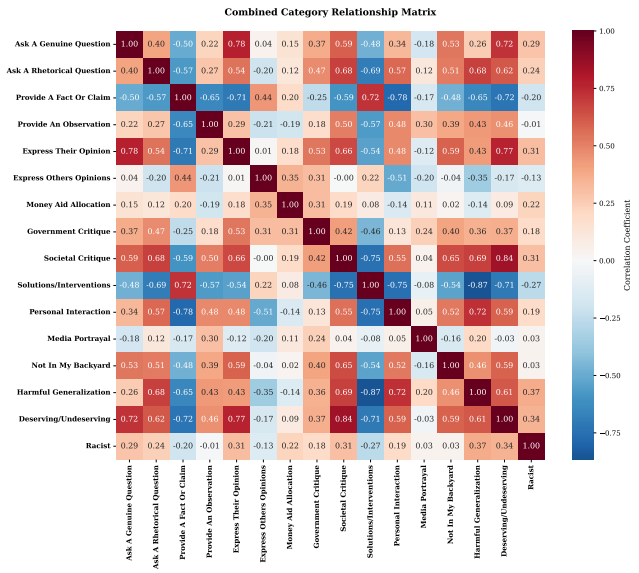

Figure 2: Correlation Matrix

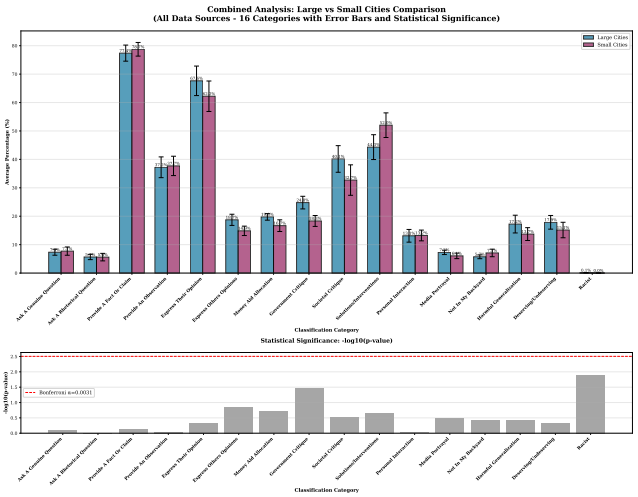

Figure 3: Large Small City Comparison

deserving" and "ask a genuine question" could potentially indicate that there is some questioning about shared beliefs that tend to blame PEHs for their fate (Sandel 2020; Desmond 2023).

Our analysis does not reveal a significant difference between small and large cities in terms of bias against PEH (Figure 3), suggesting that the levels of homelessness in big cities, such as San Francisco, do not seem to influence the homelessness bias of the local population. Wehen determining statistical signifcance, we used the Bonferroni correction for determining significance (Weisstein 2004) since this is a multi-level classification analysis. However, in Figure 4 we see that there are several significant correlations between the multimodal PEH bias classification categories and the media type. For example, meeting minutes and news sources discuss solutions/interventions more frequently than social media. Additiionally, social media posts are more likely to express harmful generalizations or opinions about the deservingness of PEH.

Ethics

The principle of beneficence, which maximizes benefits while minimizing potential harms (Beauchamp 2008), is critical to our research. It is also important to promote fairness, especially when dealing with biases towards PEH. These ethical principles are especially important in socially challenging topics such as homelessness alleviation. We have been working in close collaboration with specialized

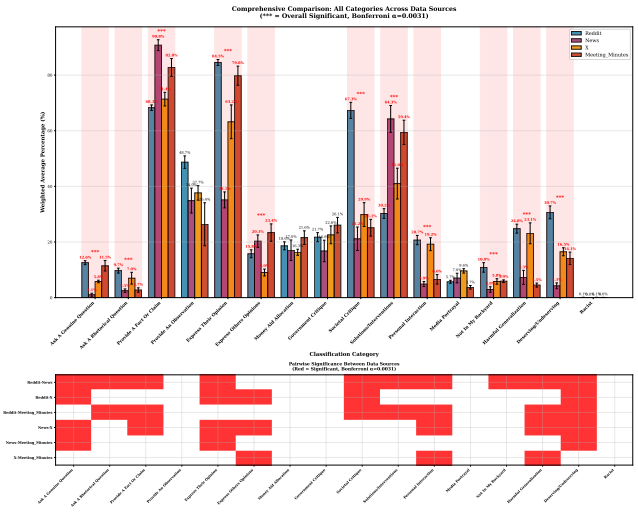

Figure 4: Data Source Comparison

non-profits in the city of South Bend to make guided decisions in the human manual annotation of homelessness biases.

We make sure that privacy is paramount. All data is anonymized to remove PII using spaCy, adhering to ethical standards for data privacy. The anonymization process ensures that individuals' identities are protected, while still allowing for valuable insights to be drawn from the data. For this process, we received IRB approval to scrape public data, and we ensure that proper guidelines and ethics are followed when using this data.

Limitations

Drawing text content from diverse sources, including Reddit, X, news sources, and city council meeting minutes, provides a robust dataset for analysis. However, it cannot be said that this dataset encompasses all of the available dialogue concerning homelessness. Future research might benefit from including novel data sources in order to capture discourse that is currently underrepresented in our final dataset. The geographic scope of our data collection is confined to 10 specific U.S. cities. Although these cities were strategically chosen to represent varying demographics and homelessness rates, they do not represent the full spectrum of socio-economic and cultural contexts across the entire United States, let alone globally.

Furthermore, despite our expanded multimodal PEH bias classification criteria and rigorous human annotation processes, the inherent complexity and subjectivity of identifying and categorizing social bias remain a challenge. Subtle, implicit biases that do not involve overt discriminatory language are difficult for automated systems to fully capture, even with advanced LLMs.

Conclusion

This research introduces a novel multi-modal dataset and demonstrates the effectiveness of Large Language Models (LLMs) in identifying and classifying homelessness bias in online and offline public discourse. Our comprehensive evaluation shows that while local LLMs may exhibit initial inconsistencies, their performance significantly improves with in-context learning, approaching the capabilities of powerful closed-source models. This highlights the potential for scalable and accessible social biases detection solutions, which are valuable tools to combat urgent social challenges (such as homelessness) by acting on public opinion. The observed variations in bias prevalence across cities and media platforms underscore the heterogeneous nature of public perception, emphasizing the necessity for context-specific interventions when aiming to alleviate homelessness through acting on the social fabric.

This work aims to foster public awareness, mitigate harmful biases, and inform policy, thereby enhancing the ethical application of generative AI in addressing critical social challenges. The invaluable partnership with the City of South Bend and their non-profit collaborators has been instrumental, guiding our human annotation and ensuring the practical relevance of our findings. By developing new indicators of homelessness bias, we empower cities like South Bend with data-driven tools to counter stigmatization and facilitate more equitable approaches to homelessness alleviation.

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.We also thank the University of Notre Dame for the Strategic Framework Grant that makes this work possible. Finally, we thank Emory Smith and Lina McKimson who have worked with the Lucy Family Institute for Data & Society.

References

Beauchamp, T. 2008. The principle of beneficence in applied ethics.

Blodgett, S. L.; Barocas, S.; Daumé Iii, H.; and Wallach, H. 2020. Language (technology) is power: A critical survey of" bias" in nlp. *arXiv preprint arXiv:2005.14050*.

Bolukbasi, T.; Chang, K.-W.; Zou, J. Y.; Saligrama, V.; and Kalai, A. T. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29.

Cardoso, J. R.; Pereira, L. M.; Iversen, M. D.; and Ramos, A. L. 2014. What is gold standard and what is ground truth? *Dental press journal of orthodontics*, 19: 27–30.

Chan, A.; Okolo, C. T.; Terner, Z.; and Wang, A. 2021. The Limits of Global Inclusion in AI Development. *arXiv.org*.

Chaudhary, G. 2024. Unveiling the black box: Bringing algorithmic transparency to AI. *Masaryk University Journal of Law and Technology*, 18(1): 93–122.

City and County of San Francisco. 2024. Homeless Population. San Francisco Government Website. [Accessed 22 Jan 2025].

Clifford, S.; and Piston, S. 2017. Explaining public support for counterproductive homelessness policy: The role of disgust. *Political Behavior*, 39: 503–525.

Curto, G.; Kiritchenko, S.; Fraser, K. C.; and Nejadgholi, I. 2024. 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)*, 32–45.

de Sousa, T.; and Henry, M. 2024. The 2024 Annual Homelessness Assessment Report (AHAR) to Congress. Technical report, The U.S. Department of HUD.

Desmond, M. 2023. *Poverty, by America*. Crown.

Fornaciari, T.; Uma, A.; Paun, S.; Plank, B.; Hovy, D.; Poesio, M.; et al. 2021. Beyond black & white: Leveraging annotator disagreement via soft-label multi-task learning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics.

Gilardi, F.; Alizadeh, M.; and Kubli, M. 2023. ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30): e2305016120.

Halterman, A.; and Keith, K. A. 2024. Codebook LLMs: adapting political science codebooks for LLM use and adapting LLMs to follow codebooks. *arXiv e-prints*, arXiv–2407.

Hartvigsen, T.; Gabriel, S.; Palangi, H.; Sap, M.; Ray, D.; and Kamar, E. 2022. Toxigen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. *arXiv preprint arXiv:2203.09509*.

Honnibal, M.; Montani, I.; Van Landeghem, S.; and Boyd, A. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Jin, Z.; and Mihalcea, R. 2022. Natural language processing for policymaking. In *Handbook of computational social science for policy*, 141–162. Springer International Publishing Cham.

Karr, J.; Smith, E.; Hauenstein, M.; Curto, G.; and Chawla, N. 2025. What is Behind Homelessness Bias? Using LLMs and NLP to Mitigate Homelessness by Acting on Social Stigma. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence (IJCAI-25)*, to appear. Accepted.

Kiritchenko, S.; Rex, G. C.; Nejadgholi, I.; and Fraser, K. C. 2023. Aporophobia: An overlooked type of toxic language targeting the poor. In *The 7th Workshop on Online Abuse and Harms (WOAH)*, 113–125.

Kogan, D. 2023. pydeidentify: A Python package for deidentification of structured data. https://github.com/dtkogan/pydeidentify.

Li, M.; Chen, H.; Wang, Y.; Zhu, T.; Zhang, W.; Zhu, K.; Wong, K.-F.; and Wang, J. 2025. Understanding and mitigating the bias inheritance in llm-based data augmentation on downstream tasks. *arXiv preprint arXiv:2502.04419*.

Li, Z.; Su, Y.; Wang, H.; and Zhao, W. 2024. BuildingView: Constructing Urban Building Exteriors Databases with Street View Imagery and Multimodal Large Language Mode. *arXiv preprint arXiv:2409.19527*.

Liang, P.; Bommasani, R.; Lee, T.; Tsipras, D.; Soylu, D.; Yasunaga, M.; Zhang, Y.; Narayanan, D.; Wu, Y.; Kumar, A.; et al. 2022. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*.

Liberty, E.; Lang, K.; and Shmakov, K. 2016. Stratified sampling meets machine learning. In *International conference on machine learning*, 2320–2329. PMLR.

Lu, Q.; Zhu, L.; Xu, X.; Whittle, J.; Zowghi, D.; and Jacquet, A. 2024. Responsible AI pattern catalogue: A collection of best practices for AI governance and engineering. *ACM Computing Surveys*, 56(7): 1–35.

Matarazzo, A.; and Torlone, R. 2025. A survey on large language models with some insights on their capabilities and limitations. *arXiv preprint arXiv:2501.04040*.

Mislove, A.; Lehmann, S.; Ahn, Y.-Y.; Onnela, J.-P.; and Rosenquist, J. 2011. Understanding the Demographics of Twitter Users. *Proceedings of the Fifth International AAAI Conference on WEblogs and Social Media*.

Morales, S.; Clarisó, R.; and Cabot, J. 2024. LangBiTe: A Platform for Testing Bias in Large Language Models. *arXiv preprint arXiv:2404.18558*.

Nadeem, M.; Bethke, A.; and Reddy, S. 2020. StereoSet: Measuring stereotypical bias in pretrained language models. *arXiv preprint arXiv:2004.09456*.

OECD. 2024. OECD Toolkit to Combat Homelessness. OECD, Paris. [Accessed 21 Jan 2025].

Opitz, J.; and Burst, S. 2019. Macro f1 and macro f1. *arXiv preprint arXiv:1911.03347*.

Ranjit, J.; Joshi, B.; Dorn, R.; Petry, L.; Koumoundouros, O.; Bottarini, J.; Liu, P.; Rice, E.; and Swayamdipta, S. 2024. OATH-Frames: Characterizing Online Attitudes Towards Homelessness with LLM Assistants. *arXiv preprint arXiv:2406.14883*.

Rex, G. C.; Kiritchenko, S.; Siddiqui, M. H. F.; Nejadgholi, I.; and Fraser, K. C. 2025. 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*.

Sandel, M. J. 2020. *The tyranny of merit*. Penguin Random House.

Srivastava, A.; Rastogi, A.; Rao, A.; Shoeb, A. A.; Abid, A.; Fisch, A.; Brown, A. R.; Santoro, A.; Gupta, A.; Garriga-Alonso, A.; et al. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on machine learning research*.

UNESCO. 2021. Recommendation on the Ethics of Artificial Intelligence.

Union, E. 2024. Regulation (EU) 673 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). *Official Journal of the European Union*.

Wang, A.; Singh, A.; Michael, J.; Hill, F.; Levy, O.; and Bowman, S. R. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.

Wang, Y.; Yao, Q.; Kwok, J. T.; and Ni, L. M. 2020. Generalizing from a few examples: A survey on few-shot learning. *ACM computing surveys (csur)*, 53(3): 1–34.

Weisstein, E. W. 2004. Bonferroni correction. *https://mathworld. wolfram. com/*.

Appendix

Category	Reddit		News		Meeting Minutes		X (Twitter)	
	Zero	Few	Zero	Few	Zero	Few	Zero	Few
Ask Genuine Question	40.22	**61.55**	51.83	**72.03**	61.02	**94.40**	**78.35**	71.12
Ask Rhetorical Question	45.16	**79.00**	84.87	**91.09**	59.20	**63.27**	**65.15**	62.27
Provide Fact/Claim	74.83	**84.83**	49.57	**87.55**	58.68	**59.08**	60.42	**62.34**
Provide Observation	46.36	**76.05**	43.72	**90.64**	**72.32**	50.92	64.02	**80.66**
Express Opinion	43.62	**77.71**	**66.22**	52.22	58.21	**91.98**	86.51	**89.33**
Express Others Opinions	44.73	**62.32**	41.88	**71.19**	86.02	**90.34**	75.39	**80.72**
Money Aid Allocation	44.33	**72.74**	**88.05**	70.87	58.89	**65.82**	64.23	**79.06**
Government Critique	52.38	**67.33**	57.15	**65.55**	56.66	**63.68**	**85.51**	52.98
Societal Critique	54.37	**73.14**	**88.49**	56.65	**89.94**	75.00	59.54	**78.87**
Solutions/Interventions	**83.69**	62.70	79.59	**86.45**	**72.46**	64.91	44.91	**60.00**
Personal Interaction	76.06	**77.54**	46.11	**63.57**	46.58	**55.52**	46.49	**59.30**
Media Portrayal	66.04	**91.60**	77.19	**78.09**	**69.68**	60.30	**76.03**	61.17
Not in My Backyard	55.50	**84.86**	**81.35**	65.47	**68.69**	67.87	80.22	**81.12**
Harmful Generalization	88.01	**91.90**	82.13	**90.43**	56.50	**59.20**	46.68	**92.26**
Deserving/Undeserving	**75.33**	62.56	70.28	**93.61**	68.54	**89.77**	**62.01**	53.01
Racist	55.75	**68.97**	80.75	**89.28**	40.48	**89.50**	78.42	**78.96**

Table 5: Category-wise F1 Scores for GEMINI Model

Category	Reddit		News		Meeting Minutes		X (Twitter)	
	Zero	Few	Zero	Few	Zero	Few	Zero	Few
Ask Genuine Question	**16.09**	13.95	**10.53**	9.52	0.00	0.00	**37.29**	15.87
Ask Rhetorical Question	**27.10**	26.42	12.90	**23.68**	0.00	**33.33**	10.53	**26.09**
Deserving/Undeserving	**4.26**	3.92	0.00	**7.41**	0.00	**50.00**	**25.00**	0.00
Express Others Opinions	**21.67**	15.38	5.13	**5.56**	**18.82**	8.57	6.25	6.25
Express Opinion	55.18	**62.89**	40.12	**44.27**	16.98	**26.36**	45.62	**56.10**
Government Critique	**20.31**	20.00	**19.80**	19.67	9.84	**10.34**	15.15	**24.69**
Harmful Generalization	17.24	**20.59**	**13.56**	11.90	0.00	0.00	0.00	**9.09**
Media Portrayal	0.00	0.00	0.00	**16.00**	0.00	**20.00**	0.00	0.00
Money Aid Allocation	11.11	**22.22**	20.37	**25.56**	**32.05**	17.27	29.63	**32.00**
Not in My Backyard	**23.19**	7.59	**16.67**	0.00	0.00	0.00	0.00	**19.05**
Personal Interaction	16.49	**18.87**	0.00	**7.02**	14.29	**26.67**	**17.39**	14.81
Provide Fact/Claim	49.16	**55.68**	61.59	**71.35**	69.04	**73.59**	60.24	**76.74**
Provide Observation	28.98	**36.91**	**4.40**	4.11	**6.00**	3.70	**7.23**	4.55
Racist	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Societal Critique	**22.10**	20.10	0.00	**20.62**	**10.00**	4.76	8.00	**16.00**
Solutions/Interventions	27.31	**32.51**	35.47	**41.97**	44.94	**53.26**	43.69	**51.38**

Table 6: Category-wise F1 Scores for GROK Model

Category	Reddit		News		Meeting Minutes		X (Twitter)	
	Zero	Few	Zero	Few	Zero	Few	Zero	Few
Ask Genuine Question	12.90	**47.37**	0.00	**21.95**	28.57	28.57	**44.90**	37.93
Ask Rhetorical Question	2.08	**14.68**	4.76	**27.07**	0.00	**22.22**	0.00	0.00
Deserving/Undeserving	**5.97**	4.11	4.44	**5.68**	**22.22**	4.55	0.00	**3.64**
Express Others Opinions	20.51	**29.93**	0.00	**11.11**	**10.00**	9.09	0.00	**4.55**
Express Opinion	**86.10**	66.13	**76.19**	71.78	**37.13**	29.15	66.35	**75.56**
Government Critique	**42.86**	32.00	**36.36**	34.48	**23.53**	16.39	7.84	**28.57**
Harmful Generalization	**31.53**	28.33	**35.94**	19.05	**11.43**	8.60	**13.33**	7.50
Media Portrayal	0.00	**1.80**	0.00	**4.69**	0.00	0.00	0.00	**1.77**
Money Aid Allocation	3.03	**17.20**	5.06	**42.01**	0.00	**30.46**	3.03	**50.82**
Not in My Backyard	**21.77**	13.33	**13.73**	10.00	0.00	0.00	8.16	**26.67**
Personal Interaction	3.64	**15.15**	0.00	**8.82**	0.00	0.00	0.00	**15.38**
Provide Fact/Claim	26.11	**66.67**	36.67	**62.03**	**67.27**	18.09	**53.61**	16.67
Provide Observation	14.18	**38.68**	**16.00**	7.06	**15.38**	8.70	**7.41**	0.00
Racist	0.00	**1.71**	0.00	**6.25**	0.00	0.00	0.00	0.00
Societal Critique	**38.87**	35.33	**29.49**	18.38	**18.18**	14.91	**15.62**	10.87
Solutions/Interventions	1.12	**62.14**	6.49	**57.69**	7.07	**64.57**	1.16	**64.79**

Table 7: Category-wise F1 Scores for LLAMA Model

Category	Reddit		News		Meeting Minutes		X (Twitter)	
	Zero	Few	Zero	Few	Zero	Few	Zero	Few
Ask Genuine Question	15.87	**43.24**	12.50	**22.22**	25.00	25.00	**48.98**	36.36
Ask Rhetorical Question	**5.83**	4.26	**13.64**	9.52	**22.22**	0.00	0.00	0.00
Deserving/Undeserving	5.56	**10.91**	0.00	**11.68**	0.00	0.00	**50.00**	0.00
Express Others Opinions	0.00	**6.90**	6.90	**7.41**	0.00	0.00	0.00	0.00
Express Opinion	**74.58**	66.67	61.87	**70.33**	34.88	**45.56**	50.64	**56.50**
Government Critique	**36.96**	35.74	22.50	**37.50**	24.24	**34.86**	12.77	**33.51**
Harmful Generalization	30.48	**38.89**	25.71	**27.45**	25.00	25.00	0.00	**16.33**
Media Portrayal	**2.63**	0.00	1.89	**7.27**	2.70	**3.33**	0.00	0.00
Money Aid Allocation	0.00	**9.88**	14.46	**36.84**	**21.85**	1.92	23.08	**25.29**
Not in My Backyard	**25.00**	24.20	5.08	**8.25**	0.00	0.00	11.43	**15.38**
Personal Interaction	**12.12**	10.53	0.00	0.00	33.33	**40.00**	30.00	**37.50**
Provide Fact/Claim	12.34	**56.82**	5.87	**46.54**	34.10	**73.85**	25.81	**38.16**
Provide Observation	1.65	**37.16**	0.00	**5.71**	0.00	**12.50**	**15.38**	0.00
Racist	0.00	0.00	0.00	**40.00**	0.00	0.00	0.00	0.00
Societal Critique	2.22	**31.63**	0.00	**26.28**	**20.00**	15.79	0.00	**8.40**
Solutions/Interventions	**20.00**	8.60	20.00	**68.65**	30.36	**63.43**	21.89	**65.45**

Table 8: Category-wise F1 Scores for PHI4 Model

Category	Reddit		News		Meeting Minutes		X (Twitter)	
	Zero	Few	Zero	Few	Zero	Few	Zero	Few
Ask Genuine Question	57.73	**62.14**	37.21	**41.18**	**40.00**	36.36	**60.00**	59.15
Ask Rhetorical Question	**14.52**	10.20	36.62	**50.98**	**11.76**	0.00	0.00	**15.38**
Deserving/Undeserving	6.58	**21.88**	3.70	**8.45**	1.09	**6.25**	2.60	**20.00**
Express Others Opinions	**25.43**	9.84	12.56	**17.02**	**16.84**	16.00	0.00	**4.35**
Express Opinion	**88.12**	74.29	69.49	**75.82**	35.37	**37.84**	75.69	**76.00**
Government Critique	**43.30**	42.75	**47.46**	45.45	20.09	**30.16**	**43.31**	38.86
Harmful Generalization	30.81	**42.86**	29.81	**35.24**	3.92	**6.67**	8.40	**11.32**
Media Portrayal	**4.00**	3.28	**1.27**	0.00	1.01	**3.39**	2.60	**2.90**
Money Aid Allocation	40.28	**43.21**	41.51	**48.94**	57.03	**62.20**	**59.86**	51.22
Not in My Backyard	17.59	**34.85**	8.38	**20.20**	0.00	0.00	3.87	**10.17**
Personal Interaction	35.15	**35.94**	9.72	**14.18**	3.47	**19.51**	17.14	**23.33**
Provide Fact/Claim	65.64	**71.45**	59.52	**83.66**	86.33	**86.81**	71.81	**80.94**
Provide Observation	45.48	**54.44**	**6.99**	5.00	3.68	**7.84**	**6.12**	0.00
Racist	0.00	0.00	0.00	**28.57**	0.00	0.00	0.00	0.00
Societal Critique	35.00	**38.99**	22.67	**24.49**	12.90	**13.79**	7.59	**19.35**
Solutions/Interventions	60.37	**65.67**	61.07	**67.99**	68.78	**75.39**	70.09	**72.21**

Table 9: Category-wise F1 Scores for QWEN Model

Category	Reddit	News	Meeting Minutes	X (Twitter)
Racist	40.00	0.00	0.00	0.00
Ask Genuine Question	21.05	0.00	0.00	0.00
Ask Rhetorical Question	0.00	0.00	0.00	0.00
Deserving/Undeserving	20.00	0.00	0.00	0.00
Express Others Opinions	41.46	0.00	30.00	0.00
Express Opinion	95.38	89.70	0.00	87.88
Government Critique	11.11	0.00	0.00	0.00
Harmful Generalization	12.77	0.00	0.00	0.00
Media Portrayal	40.00	0.00	0.00	0.00
Money Aid Allocation	7.55	0.00	64.91	65.67
Not in My Backyard	12.50	0.00	0.00	0.00
Personal Interaction	36.84	0.00	0.00	0.00
Provide Fact/Claim	90.91	99.45	99.35	99.32
Provide Observation	68.97	0.00	0.00	0.00
Societal Critique	61.64	0.00	0.00	0.00
Solutions/Interventions	80.70	71.83	85.93	90.37

Table 10: Category-wise F1 Scores for BERT Fine-tuned Model

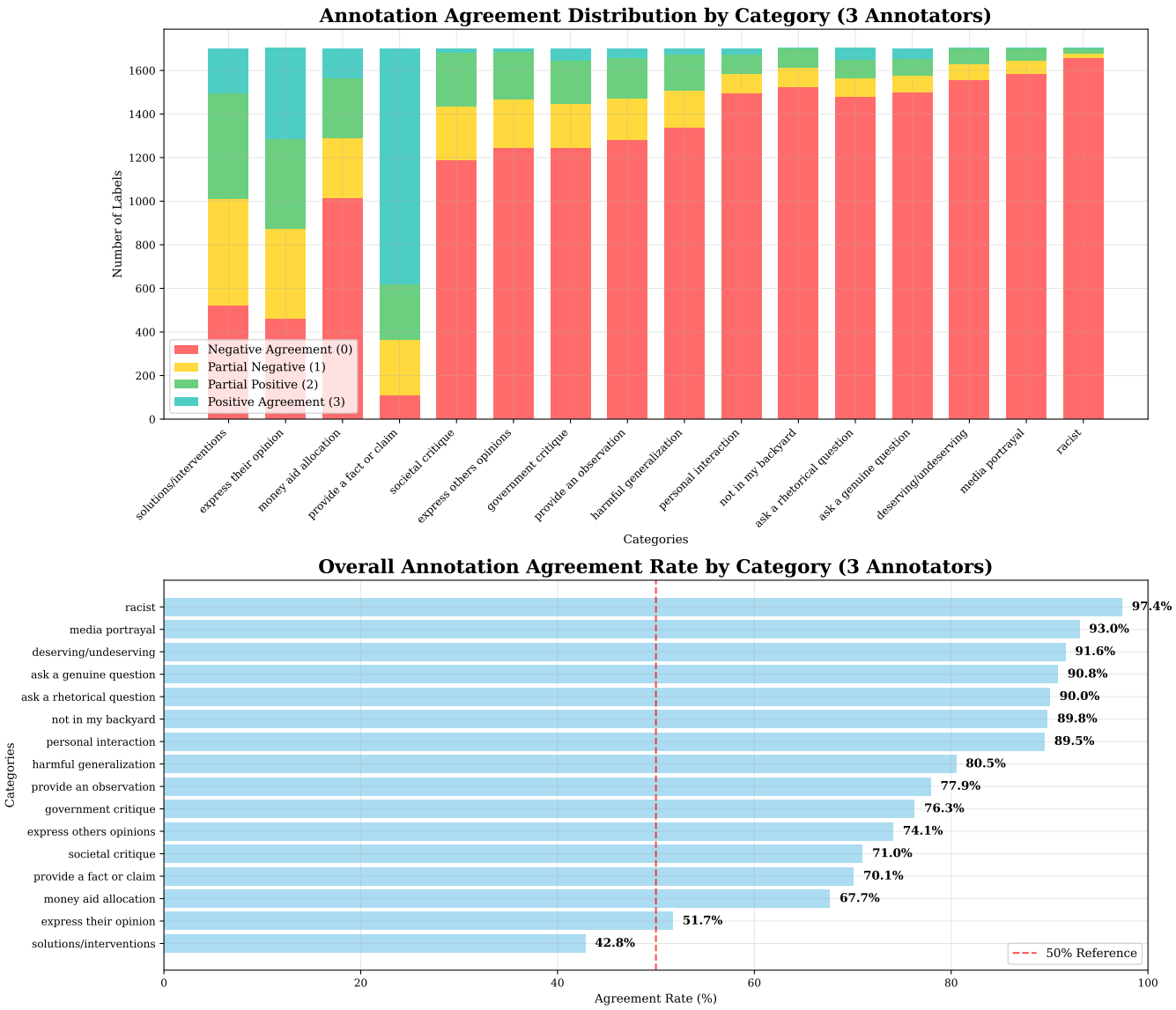

Figure 5: Annotator Agreement by Category

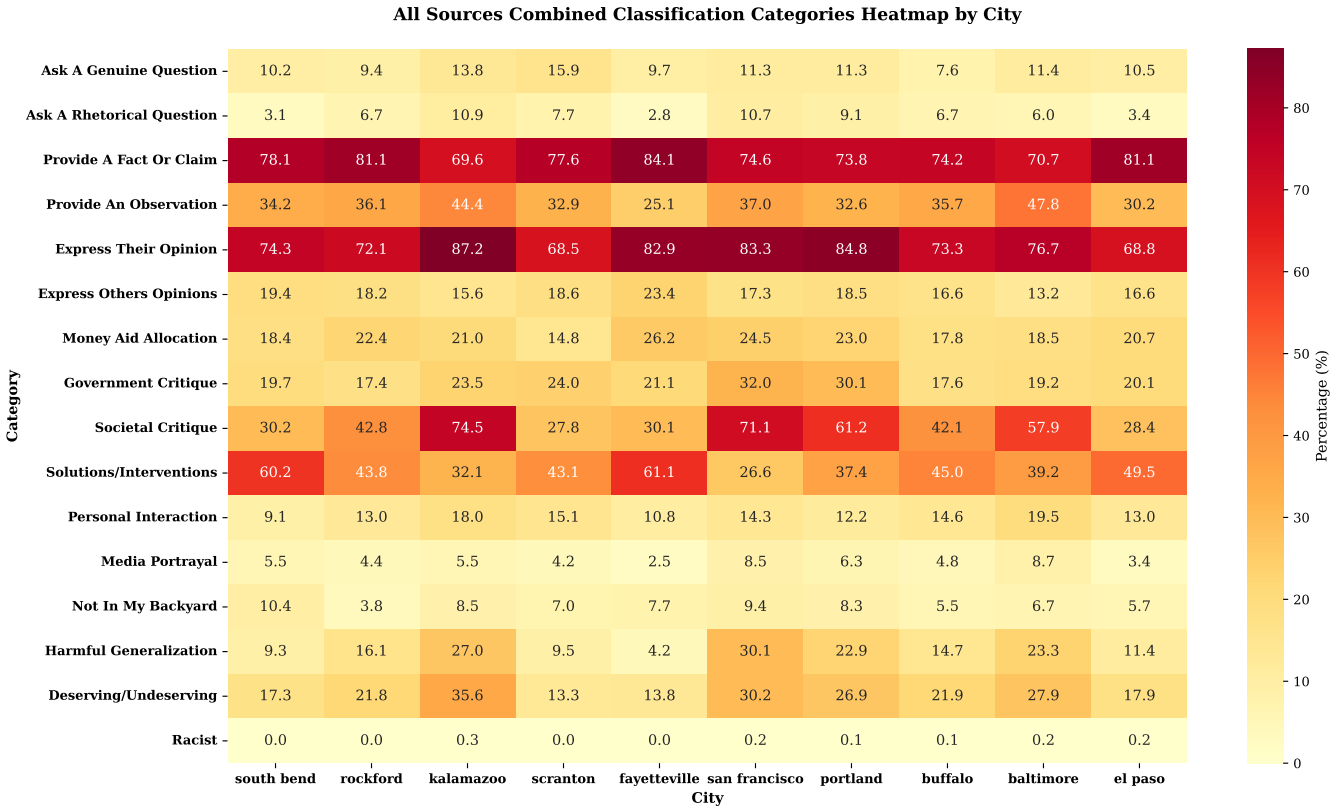

Figure 6: City Classification Heatmap

Grouping	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	1,032	131	98	0.52
2	Multnomah County, Oregon (Portland)	0.56	808,098	1,198	237	91	0.47
3	Erie County, New York (Buffalo)	0.47	951,232	1,342	134	60	0.46
4	Baltimore County, Maryland (Baltimore)	0.63	850,737	997	99	7	0.46
5	El Paso County, Texas (El Paso)	0.69	863,832	1,919	99	11	0.47
	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	Winnebago County, Illinois (Rockford)	0.57	284,591	1,583	134	29	0.45
C	Kalamazoo County, Michigan (Kalamazoo)	0.43	261,426	1297	83	25	0.46
D	Lackawanna County, Pennsylvania (Scranton)	0.38	215,672	1252	238	8	0.46
E	Washington County, Arkansas (Fayetteville)	0.60	247,331	1,466	80	32.14	0.48

*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 11: Table of US counties, used in the dataset. Counties are similar to San Francisco County, CA, and St. Joseph County, IN. The counties are similar in the rate of people below the poverty line, the rate of people with public assistance, the homelessness rate, and GINI, yet differ in racial fractionalization.