Local LLM Zero-Shot Analysis of Homelessness Discourse on Reddit

Anonymous ACL submission

Abstract

Homelessness is a persistent issue, impacting millions worldwide, and 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 policy. Online discourse on platforms such as Reddit shape public opinion. To address this, the project leverages natural language processing and large language models (LLMs) to mitigate bias against people experiencing homelessness (PEH) in online spaces. The goal is to promote awareness, reduce harmful biases, inform policy, and improve the fairness of generative AI. We gather Reddit data for 10 U.S. cities, then perform zero-shot classification, and finally, mitigation using Llama 3.2 Instruct and Qwen 2.5 7B Instruct models. The initial results highlighted the differing classifications between models and indicated that many mitigated outputs remained biased. This suggests the need for potential model refinement for the mitigation of text related to PEH.

1 Introduction

Homelessness is a persistent issue that affects millions of people worldwide. In the United States, over 770,000 people were recorded as experiencing homelessness in 2024, the highest number ever recorded (de Sousa and Henry, 2024). Although structural causes of homelessness have garnered attention, the social stigmatization of the issue remains a significant barrier to alleviating homelessness. Bias caused by such stigmatization shifts public opinion regarding the issue and contributes to marginalizing and dehumanizing those affected. This shift in public perception of homelessness leads to a shift in voting and, thus, policy aimed at addressing the issue (Clifford and Piston, 2017).

Discourse on social media platforms such as Reddit can influence the perceptions and opinions of users greatly, and these AI tools provide the opportunity to mitigate the harmful effects of biased and misleading discourse on the homeless population. Although the project only focuses on online textual discourse, it can have real-world implications by shaping public perceptions and influencing policy discussions related to homelessness. Additionally, the project can serve as a foundation for future work related to the intersection of artificial intelligence, social media, and public opinion.

To address the pervasive stigma and bias toward homelessness, we leverage natural language processing (NLP) and large language models (LLMs) to address biases against those experiencing homelessness in online spaces. We present the following research questions (RQs):

RQ1: What are the biases of homelessness discourse on Reddit?

RQ2: How well do local large language models perform zero-shot bias classification of English textual discourse about homelessness?

RQ3: How well do local LLMs mitigate biases for online English textual discourse?

To solve these RQs, we do the following tasks.

(1) We collect data from Reddit on homelessness discourse between 2015 and 2025 for 10 U.S. cities using the PEH lexicon (Karr et al., 2025).

(2) We anonymize the data using spaCy to preserve anonymity.

(3) We classify the Reddit biases towards PEH with the OATH-Frames (Ranjit et al., 2024) by using Llama 3.2 3B Instruct, Qwen 2.5 7B Instruct, and human annotators.

(4) Finally we mitigate the data using Llama 3.2 3B Instruct and Qwen 2.5 7B Instruct and then reclassify the mitigated results with the LLMs.

Our approach aims to foster greater public awareness, reduce the spread of harmful biases, informing policy, and improving the reliability and fairness of generative AI models in the topic of homelessness. However, we recognize the potential risks associated with relying on AI to identify bias in online discourse. If the AI is incorrectly missing homelessness bias or falsely flagging non-biased content, people may be misled. Therefore, this project is guided by the principle of beneficence, which maximizes benefits while minimizing potential harms (Beauchamp, 2008).

2 Related Work

2.1 Current Homelessness Bias Classification Techniques

Previous studies have used LLMs to classify and analyze online content that is considered biased against the poor (Kiritchenko et al., 2023; Curto et al., 2024; Rex et al., 2025). This has been done by searching through 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 people experiencing homelessness (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).

OATH (Ranjit et al., 2024) has one of the most comprehensive pipelines for homelessness bias classification. The OATH-Frame categorizes biases into a variety of predicted frames for critiques, responses, and perceptions, such as 'government critique', 'not in my backyard', and 'harmful generalization'.

2.2 Current Mitigation Techniques

Efforts to mitigate bias in machine learning include several strategies. Although few mitigation techniques have been applied to the homelessness domain, several standard approaches have been applied to adjacent domains. One prominent approach is re-weighting or re-sampling the training data to balance representation across demographic groups (Kamiran and Calders, 2012; Gallegos et al., 2024). Another technique is adversarial debiasing, where a secondary model is trained to remove bias from the primary model's predictions (Zhang et al., 2018). These techniques have been successfully applied in domains such as criminal justice (Hardt et al., 2016).

For NLP applications, counterfactual data augmentation and bias-controlled fine-tuning have been used to improve fairness in text classification tasks (Feng et al., 2021; Dinan et al., 2019). Additionally, multi-agent LLM approaches have been developed to reduce bias (Borah and Mihalcea, 2024). Interpretability methods like SHAP and LIME can reveal which features contribute to biased predictions, enabling targeted mitigation (Lundberg and Lee, 2017; Ribeiro et al., 2016).

3 Methodology

As noted in Figure 1, we collect data from Reddit by using the PEH lexicon for scraping (Karr et al., 2025). Then we anonymize the data with spaCy (Honnibal et al., 2020) to remove personal identifiable information (PII). Then we classify the comments' biases using OATH-Frames (Ranjit et al., 2024). For classification, we use both human annotators and Llama 3.2 3B Instruct and Qwen 2.5 7B Instruct LLMs. Finally, we mitigate the data using Llama 3.2 3B Instruct and Qwen 2.5 7B Instruct and reclassify the data to see how well local models can mitigate English textual online discourse.

3.1 Data Collection

We collected English Reddit data from 10 cities across the U.S. as documented in prior work (Karr et al., 2025). They chose five cities similar to South Bend, Indiana, USA, and five similar to San Francisco, California, USA. In order to collect a substantial amount of data, we ensured that all of the cities had at minimum 50 Reddit posts between January 1st, 2015, and January 1st, 2025. If one of their cities had fewer than 50 comments, we replaced it with another city that was in its list of 20 k-Nearest-Neighbors (kNNs). The groups were counties that the cities were in and were grouped by the following statistics: RPP (Rate of People Below Poverty Line), RPA (Rate of People With Public Assistance), Homelessness Rate, and GINI (Income Inequality). To collect this data, we scraped Reddit posts and comments that were part of the PEH lexicon (Karr et al., 2025), which includes words such as 'homeless', 'unhoused', and 'beggar.'

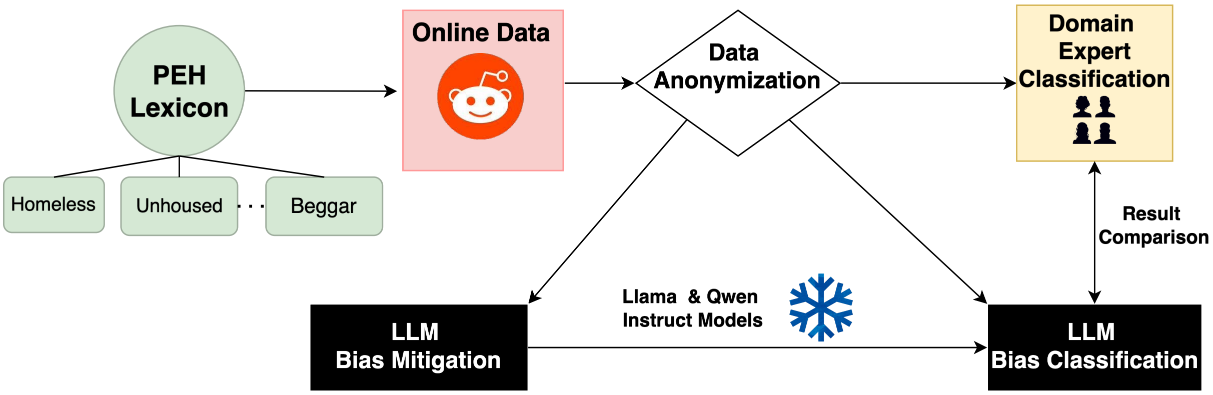

Figure 1: We collect Reddit Data on homelessness discourse using a prior lexicon. Then we anonymize the data and have both LLMs and domain experts classify the data to determine reliability. Finally, LLMs mitigate the data in order to reduce biases.

3.2 Data Anonymization

Prioritizing the anonymization of Reddit data is essential for research and privacy protection. We leveraged the capabilities of the spaCy natural language processing library (Honnibal et al., 2020). This technique allowed us to automatically identify and mask potentially Personally Identifiable Information (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.

This multi-faceted anonymization strategy was crucial in establishing a dataset that respects user privacy while retaining the linguistic characteristics essential for our analysis of bias and the development of mitigation techniques.

3.3 PEH Bias Classification

We created a bias classification for homelessness discourse based on a combination of prior work (Rex et al., 2025; Ranjit et al., 2024).

Comment Type: is classified as either 'direct' or 'reporting' (Rex et al., 2025). This original taxonomy was on bias against the poor, or aporophobia, so we adapted it to PEH. The definitions that we use are the following:
Direct- The speaker expresses their own views about PEH
Reporting - The speaker describes or criticizes others' views/behaviors regarding PEH.

We are also using OATH-Frames (Ranjit et al., 2024), which is an existing bias classification for PEH. The group of categories are critique, response, and perception, and the definitions for the categories can be found in their paper. The categories have multiple terms, and based on the Reddit post, we can classify a post as having a variety of terms, based on the biases. The category in each group are as follows:
Critique Categories - 'money aid allocation', 'government critique', and 'societal critique.'
Response Categories - 'solutions/interventions.'
Perception Types - 'personal interaction', 'media portrayal', 'not in my backyard', 'harmful generalization', and 'deserving/undeserving.'

Finally, the model is asked to explicitly identify if the comment contains **racist** content. We include this since prior political science works states that racial fractionalization influences homelessness bias (Alesina and Glaeser, 2013). Therefore, we see if racist remarks are prevalent in homelessness discourse by labeling each post as racist or not.

3.4 Gold Standard & Soft Labeling

We created a gold standard (Cardoso et al., 2014) that had 50 Reddit comments from each of the 10 cities, for a total of 500. This form of stratified sampling is known as equal representation (Liberty et al., 2016), which improves accuracy when strata from cities differs significantly. Given that we have five small cities similar to South Bend and five large cities similar to San Francisco, the strata between the number of comments between large and small cities will vary.

We had two human annotators classify the data who are familiar with PEH. Given that biases vary from person to person, it is expected that labeling will differ slightly. Therefore, we utilize soft labeling (Fornaciari et al., 2021), which takes an average of annotators responses. Soft labeling is effective when there is disagreement, since it can be challenging to determine what is biased or not in certain instances.

3.5 Model Selection

The core of our bias analysis and mitigation pipeline relies on the capabilities of an LLM. After consideration of various options, we selected the Llama 3.2 3B Instruct and Qwen 2.5 7B Instruct models for this purpose. Our decision was driven by the following key factors:

Local Deployment and Cost Efficiency: A significant advantage of the models is that they are open-source nature, allowing for local deployment without the need for costly API access and per-token charges associated with proprietary models. This was a crucial consideration given the resource constraints of our project.

Balance of Size and Performance: The three and seven billion parameter size of the models represents a favorable trade-off between model complexity and the computational resources required for local operation. While larger models might offer superior performance in some tasks, their demanding hardware requirements can be a limiting factor for local execution.

Suitability of the Instruct Finetuning: Initial experiments using the base version of Llama 3.2 3B for our bias classification task resulted in the model not being able to formulate answers to questions. We observed that the "Instruct" fine-tuned variant, specifically trained to follow natural language instructions and engage in dialogue-like interactions, demonstrated a markedly improved ability to understand the nuances of our prompts and provide accurate classifications. The versions, readily accessible through Hugging Face (AI, 2024; Cloud, 2024), proved to be significantly more adept at the complexities of identifying and responding to biased language.

Zero-Shot on our Data: While the instruct models are fine-tuned on answering instructions, these models are not fine-tuned on our data, nor due we fine-tune it after downloading the model. By seeing the zero-shot performance (Kojima et al., 2022) of these models, we can see how current local LLMs performs on bias related to PEH. Furthermore, we treat each prompt independently and do not chain them together to ensure fair output.

Deterministic Model: By setting the temperature of the models to 0.1, it operates in a deterministic-like structure that allows for consistent outputs when prompting the model multiple times.

By choosing the Llama 3.2 3B Instruct and Qwen 2.5 7B Instruct models, we aimed to leverage state-of-the-art LLMs that offer a strong balance of performance, local accessibility, and instruction-following capabilities, making them well-suited for our prompt-engineered approach to addressing bias against people experiencing homelessness.

3.6 LLM Bias Classification

For LLM Bias classification, we use the same prompt for each post regardless of what model is used. The prompt includes the definitions of our PEH Bias Classification as outlined in Section 3.3. We then have it output in a list which we parse and put it into a CSV. We also have it provide reasoning for its classification. The full prompt can be found in the scripts/utils.py file of our anonymized repository https://anonymous.4open.science/r/ACLSRW25-5AB2/README.md.

3.7 LLM Bias Mitigation

For bias mitigation, we ask if the original sentence is biased. Then we ask it to remove biases or make it as least biased as possible, without losing the context of the original sentence. Finally we ask if the mitigated sentence is biased, and then we perform LLM bias classification on it to compare the results to the original sentence.

4 Results

Our results detail the process and outcomes of our (1) Data Collection, (2) Gold Standard & Soft Labeling, (3) LLM Bias Classification, and (4) LLM Bias Mitigation.

4.1 Data Collection

<table>
<tr><th colspan="6">Reddit Posts & Comments Related to PEH</th></tr>
<tr><th colspan="6">Small Cities - Similar to South Bend, IN</th></tr>
<tr><th>County</th><th>City</th><th>Total Posts</th><th>Total Comments</th><th>Total Filtered Comments</th><th>Average Filtered Comment Score</th></tr>
<tr><td>St. Joseph County, Indiana</td><td>South Bend</td><td>49</td><td>1,352</td><td>196</td><td>6.29</td></tr>
<tr><td>Winnebago County, Illinois</td><td>Rockford</td><td>12</td><td>4,139</td><td>188</td><td>5.85</td></tr>
<tr><td>Kalamazoo County, Michigan</td><td>Kalamazoo</td><td>88</td><td>11,263</td><td>1,846</td><td>5.12</td></tr>
<tr><td>Lackawanna County, Pennsylvania</td><td>Scranton</td><td>8</td><td>615</td><td>79</td><td>3.59</td></tr>
<tr><td>Washington County, Arkansas</td><td>Fayetteville</td><td>12</td><td>1,157</td><td>102</td><td>5.46</td></tr>
<tr><th colspan="6">Large Cities - Similar to San Francisco, CA</th></tr>
<tr><th>County</th><th>City</th><th>Total Posts</th><th>Total Comments</th><th>Total Filtered Comments</th><th>Average Filtered Comment Score</th></tr>
<tr><td>San Francisco, California</td><td>San Francisco</td><td>579</td><td>92,965</td><td>14,777</td><td>10.67</td></tr>
<tr><td>Multnomah County, Oregon</td><td>Portland</td><td>498</td><td>102,560</td><td>15,301</td><td>17.68</td></tr>
<tr><td>Erie County, New York</td><td>Buffalo</td><td>44</td><td>10,230</td><td>589</td><td>35.28</td></tr>
<tr><td>Baltimore County, Maryland</td><td>Baltimore</td><td>222</td><td>13,464</td><td>1,215</td><td>28.89</td></tr>
<tr><td>El Paso County, Texas</td><td>El Paso</td><td>11</td><td>1,700</td><td>154</td><td>4.62</td></tr>
</table>

Table 1: Reddit Data Collection Statistics on PEH
Key: **Total Posts** - Number of Posts with a keyword in the PEH lexicon. **Total Comments** - All comments in Total Posts. **Total Filtered Comments** - Total Comments that have a keyword in the PEH lexicon.

We compiled Reddit data from 10 different cities, 5 similar to South Bend, Indiana, USA, and five similar to San Francisco, California, USA, as outlined in prior work (Karr et al., 2025). Of the cities that they chose, four of them had fewer than 50 Reddit posts between January 1st, 2015, and January 1st, 2025. Due to the lack of data, we had to replace them with other cities. Since census data in the United States is gathered by county, we searched for four counties that had cities, 3 of which were in the same kNN grouping as South Bend, and one which needed to be from the San Francisco grouping. The results of our data gathering can be seen in Table 1.

4.2 Gold Standard & Soft Labeling

The two human annotators who classified the 500 sentences are familiar with PEH. However, their agreement rate was 80.08% which is typical given that different people have different biases, and it is difficult to determine biases in some case. By using soft labeling (Fornaciari et al., 2021), we were able to understand the agreement better. If both annotators believed that a category for a sentence was biased, it received the soft label 1 (a positive). However, if only one annotator thought so, it received the soft label. If neither annotator thought so, it received the soft label 0, a negative.

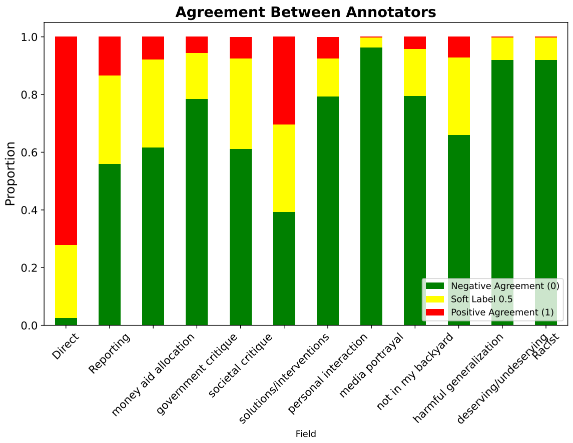

Figure 2: Agreement Between Annotators

4.3 LLM Bias Classification

As described in Section 3.6, Llama 3.2 3B Instruct and Qwen 2.5 7B Instruct classify the categories as defined by our PEH bias classification method. The confusion matrices in Figure 3 show that the classifications of Llama 3.2 3B Instruct and Qwen 2.5 7B vary widely, even though they are given the same classification prompt. This can also be seen by their low score, ranging from 0-0.31, depending on what classification category is being analyzed.

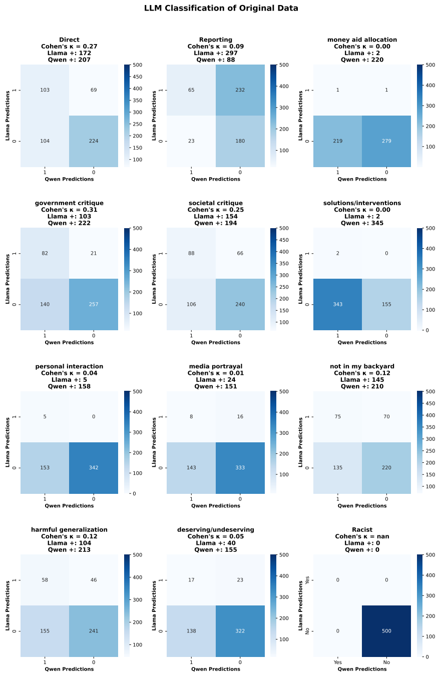

Figure 3: LLM Classification Confusion Matrices

Additionally, Figure 4 highlights the disagreement between the LLM classification and the human annotation classification. For 'direct' and 'reporting', Llama misclassifies the categories more often than Qwen. However, for the majority of the OATH-Frames, Qwen misclassified the categories more often than Llama.

4.4 LLM Bias Mitigation

As described in Section 3.7, Llama 3.2 3B Instruct and Qwen 2.5 7B Instruct determine if a Reddit post is biased, tries to mitigate, determines if the mitigated comment is biased, and then reclassifies it. Of the 500 sentences, both Llama and Qwen categorized every sentence as biased towards PEH before and after mitigation. This shows that mitigation is difficult for local LLMs. Furthermore, if you were to take all PEH bias out of a post, the post may risk loosing context.

Figure 5 highlights that posts are still biased after mitigation. In fact certain categories such as 'government critique' and 'deserving/undeserving' actually increase after Qwen mitigation.

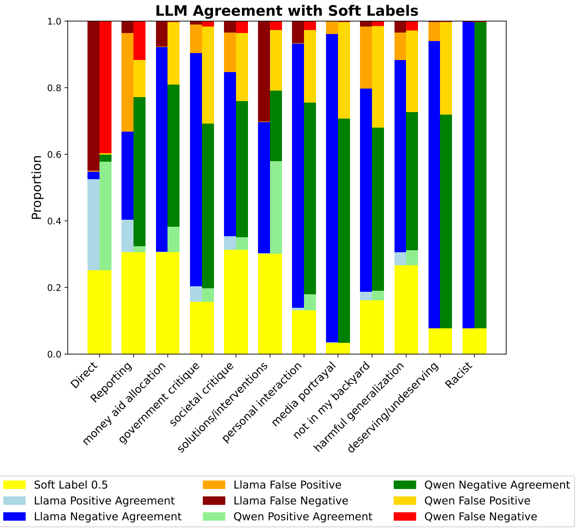

Figure 4: LLM Agreement with Human Annotators

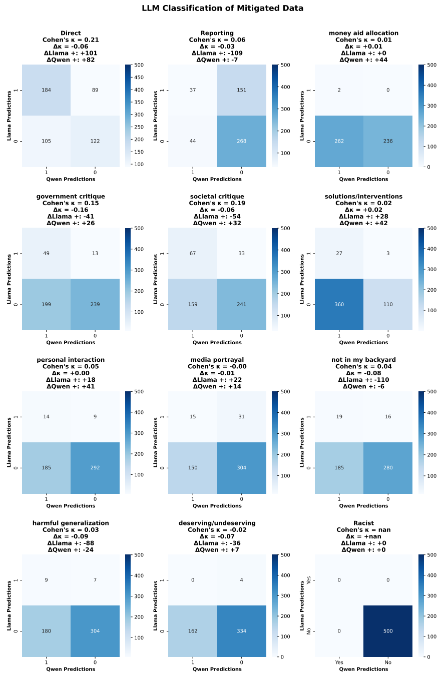

Figure 5: LLM Mitigated Confusion Matrices

5 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 . The key ethical principles guiding our methodology include the following:

Privacy and Anonymization: Ensuring privacy is paramount. All data will be anonymized to remove Personally Identifiable Information (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.

Fairness and Bias Mitigation: The central aim of this project is to mitigate bias against people experiencing homelessness. Attention was given to intersectional concerns, such as race and socioeconomic status, to prevent further marginalization of vulnerable communities. Throughout development, we evaluated and adjusted the model to ensure equitable treatment of all individuals and groups.

IRB Approval: For this project, we received IRB approval to scrape data from Reddit, and we will ensure that proper guidelines and ethics are followed when using this data.

6 Limitations

Our work is limited to small local LLMs, which may not perform as well as larger LLMs. Future work will investigate enhancing the bias classification and mitigation system through the integration of larger language models and a multi-model architecture. Larger LLMs, leveraging increased parameter counts, offer the potential for improved capture of nuanced linguistic contexts critical for accurate bias identification and mitigation. Furthermore, a multi-model approach will be examined, wherein an ensemble of LLMs with varied architectures or training objectives is combined. Additionally, it would be beneficial to use or create distinct models that specialize in textual bias (e.g., stereotyping, discriminatory language).

Since our approach is zero-shot, we do not use our gold standard as a training and testing dataset, which could improve performance. Additionally the LLM models do not mitigate the text based on the classified data, which could lead to better results.

Currently our data is limited to English Reddit textual data. APIs such as LexusNexus NewsAPI and X can be leveraged to include diverse social media, online forums, and public discourse datasets. This expanded data acquisition aims to improve the generalizability of mitigation strategies across varied online contexts and linguistic styles.

Additionally the data is limited to 10 cities in the United States. This is a subset of cities and does not represent every part of the United States, nor every part of the world. Additionally, not everyone in a city uses Reddit. Therefore, the analysis of overall biases towards PEH is very limiting.

The PEH Bias Classification categories are limiting. For example, not all OATH-Frames account for bias. For example the sentence ‘ The government should / should not use taxpayer money for people experiencing homelessness’ would be categorized as ‘money aid allocation’ regardless of the option. Sentiment analysis could be used. However, a persons' sentiment may change in long posts. This would require sentiment matching to specific parts of posts in order to be effective.

7 Conclusion

Our research represents an initial step towards leveraging LLMs for the challenging task of identifying and mitigating bias in online discourse related to homelessness by providing a Reddit dataset and doing initial testing. Our findings highlight the complexities of this issue, revealing inconsistencies in bias classification between LLMs and human annotators, as well as the difficulty LLMs face in effectively mitigating identified biases. While our results indicate that current local LLMs struggle to fully address these challenges, they also underscore the potential for AI to contribute to creating more equitable online spaces, ultimately fostering a better understanding of online textual biases that could inform improved policymaking and restore human dignity.

References

Meta AI. 2024. Llama 3.2 3b instruct. https://huggingface.co/meta-llama/Llama-3.2-3B_Instruct.

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

Tom Beauchamp. 2008. The principle of beneficence in applied ethics.

Angana Borah and Rada Mihalcea. 2024. Towards implicit bias detection and mitigation in multi-agent llm interactions. *arXiv preprint arXiv:2410.02584*.

Jefferson Rosa Cardoso, Ligia Maxwell Pereira, Maura Daly Iversen, and Adilson Luiz Ramos. 2014. What is gold standard and what is ground truth? *Dental press journal of orthodontics*, 19:27–30.

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

Alibaba Cloud. 2024. Qwen 2.5 7b instruct. https://huggingface.co/Qwen/Qwen2.5-7B-Instruct.

Georgina Curto, Svetlana Kiritchenko, Kathleen C Fraser, and Isar Nejadgholi. 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)*, pages 32–45.

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

Emily Dinan, Angela Fan, Adina Williams, Jack Urbanek, Douwe Kiela, and Jason Weston. 2019. Queens are powerful too: Mitigating gender bias in dialogue generation. *arXiv preprint arXiv:1911.03842*.

Steven Y Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard Hovy. 2021. A survey of data augmentation approaches for nlp. *arXiv preprint arXiv:2105.03075*.

Tommaso Fornaciari, Alexandra Uma, Silviu Paun, Barbara Plank, Dirk Hovy, Massimo Poesio, and 1 others. 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.

Isabel O Gallegos, Ryan A Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K Ahmed. 2024. Bias and fairness in large language models: A survey. *Computational Linguistics*, 50(3):1097–1179.

Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29.

Matthew Honnibal, Ines Montani, Sophie Van Landeghem, and Adriane Boyd. 2020. spacy: Industrial-strength natural language processing in python.

Faisal Kamiran and Toon Calders. 2012. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, 33(1):1–33.

Jonathan Karr, Emory Smith, Matthew Hauenstein, Georgina Curto, and Nitesh Chawla. 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)*, page to appear. Accepted.

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

Daniel Kogan. 2023. pydeidentify: A python package for de-identification of structured data. https://github.com/dtkogan/pydeidentify.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Edo Liberty, Kevin Lang, and Konstantin Shmakov. 2016. Stratified sampling meets machine learning. In *International conference on machine learning*, pages 2320–2329. PMLR.

Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.

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

Georgina Curto Rex, Svetlana Kiritchenko, Muhammad Hammad Fahim Siddiqui, Isar Nejadgholi, and Kathleen C Fraser. 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*.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144.

Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. 2018. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 335–340.