Can LLMs Contribute to Social Inclusion? A Zero-Shot Analysis to Automate Homelessness Bias Detection on Reddit Data

Jonathan A. Karr Jr.[1], Ben Herbst[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

Correspondence: jkarr@nd.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 policy. Online discourse on platforms such as Reddit shape public opinion. We present novel methods that builds on natural language processing (NLP) and large language models (LLMs) research to mitigate bias against people experiencing homelessness (PEH) in online spaces. We gather Reddit data for 10 U.S. cities, then perform zero-shot classification, and finally, we apply mitigation techniques using Llama 3.2 Instruct and Qwen 2.5 7B Instruct models. The results highlight the inconsistencies between LLMs when used to classify homelessness bias and the low effectiveness of GenAI tools to mitigate PEH online. The ultimate goal of this work is to promote awareness on bias against PEH, produce new indicators that inform policy, and improve the fairness of GenAI.

Code: https://github.com/Homelessness-Project/ACLSRW25 .

1 Introduction

Homelessness is a persistent social challenge that affects millions of people worldwide. The Organization for Economic Cooperation and Development (OECD) reports that 2.2 million people experience homelessness (PEH) in its 35 member 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). Recent reports inform that the number of PEH have quadrupled in San Francisco (City and County of San Francisco, 2024). In this context, there is a growing call for a shift from traditional management of homelessness (which focuses on providing material resources), to comprehensive support approaches that also tackle 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 makes it harder for policy makers to approve and implement social measures that aim to mitigate homelessness (Curto et al., 2024; Rex et al., 2025). Public perception of homelessness influences public voting in elections and, thus, it has an impact on policy aimed at addressing it (Clifford and Piston, 2017).

While data found online constitutes a non-random representation of the overall population (not all genders and identity groups are equally represented (Chan et al., 2021; Mislove et al., 2011)), it constitutes an affordable and relatively fast method to obtain preliminary indicators on social biases expressed through language. This study contributes to the nascent field of research of agentic LLMs for social impact, and we present results on the effectiveness of LLMs as classifiers of online data to generate and track new indices of homelessness bias in different US regions and correlate these indices with the actual levels of homelessness and policy making.

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

An enormous amount of research has been dedicated to identifying, tracking, and mitigating bias in AI, especially connected to the demographic balance of datasets used in Machine Learning predictions as well as in the results offered by LLMs. However, less attention has been paid to using online data as a source of human social biases and the fine-tuning of LLMs as automatic classifiers to generate new socio-economic indexes.

2.1 Bias Mitigation in Machine Learning predictions

Efforts to mitigate bias in machine learning include several strategies, and these focus on the balanced representation of demographic groups within the dataset, but not on mitigating actual social biases. One prominent approach to bias mitigation in data consists in 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).

2.2 Social Bias against PEH Classification Techniques

Previous studies have evaluated the effectiveness of LLMs as classifiers of bias against the poor in online data (Kiritchenko et al., 2023; Curto et al., 2024; Rex et al., 2025; Ranjit et al., 2024) . 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). These studies and highlight the need to conduct a deeper and specific analysis focusing on bias against PEH, with purposely created lexicons and the collection of data from a diversity of both online and offline sources.

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'.

3 Methodology

As noted in Figure 1, we collect data from Reddit by using the PEH lexicon (Karr et al., 2025).

Then we anonymize the data with spaCy (Honnibal et al., 2020) to remove personal identifiable information (PII). We identify bias against PEH and classify the types of biases using OATH-Frames (Ranjit et al., 2024). We use both human annotators and LLMs as PEH bias classifiers (namely, Llama 3.2 3B Instruct and Qwen 2.5 7B Instruct). Finally, we evaluate the effectiveness of Llama 3.2 3B Instruct and Qwen 2.5 7B Instruct to mitigate

homelessness bias in the data.

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). For that purpose, we 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.'

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.

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

3.3 PEH Bias Classification

We created a bias classification for homelessness discourse based on a combination of the categories proposed by (Rex et al., 2025) and (Ranjit et al., 2024).

First Classification Grouping: critique / response / perception: We are using the classification proposed in OATH-Frames for bias against PEH (Ranjit et al., 2024) 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.'

Second Classification Grouping: comment type: Prior work has classified com is classified as either 'direct' or 'reporting' (Rex et al., 2025). This original taxonomy was defined for bias against the poor, or aporophobia, and we adapted it to bias against PEH. We reframed this into Express Their Opinion and Others' Opinion. In addition we added Provide a Fact or Claim and Provide an Observation. Finally, we added Ask a Genuine Question and Ask a Rhetorical Question since questions are common on Reddit.

Finally, we explicitly identify if the comment contains **racist** content to provide insights regarding the potential correlation between racial fractionalization and the public support towards policies to mitigate poverty and homelessness (Alesina and Glaeser, 2013).

3.4 Manually Annotated Baseline

We create a manually annotated baseline (Cardoso et al., 2014) that contains 50 Reddit comments from each of the 10 cities (a total of 500 Reddit comments). 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. We selected Llama 3.2 3B Instruct and Qwen 2.5 7B Instruct models based on the following factors:

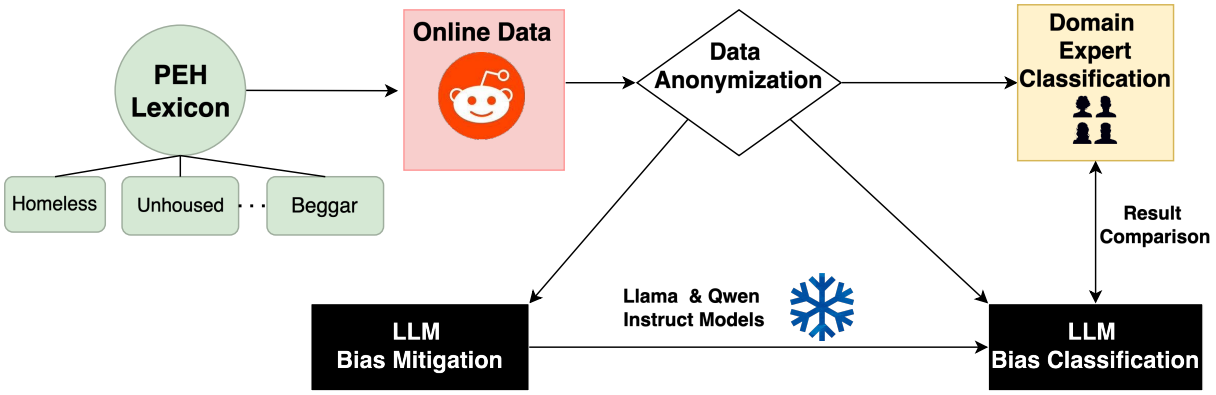

Figure 1: We collect Reddit Data on homelessness discourse using a prior lexicon. Then we anonymize the data and have both LLMs and domain experts classify the data to determine reliability. Finally, LLMs mitigate the data in order to reduce biases.

Local Deployment and Cost Efficiency: They are open-source, allowing for local deployment without the need for costly API access and per-token charges associated with proprietary models.

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 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 repository.

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 include (1) Data Collection, (2) Gold Standard & Soft Labeling, (3) LLM Bias Classification, and (4) LLM Bias Mitigation.

4.1 Data Collection

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

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.

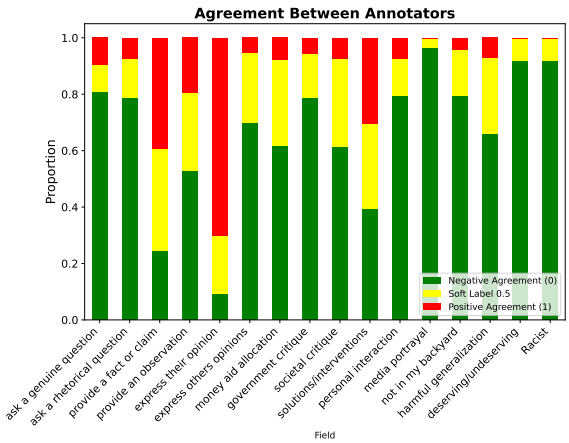

Figure 2: Agreement Between Annotators

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 4 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.

Additionally, Figure 3 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.

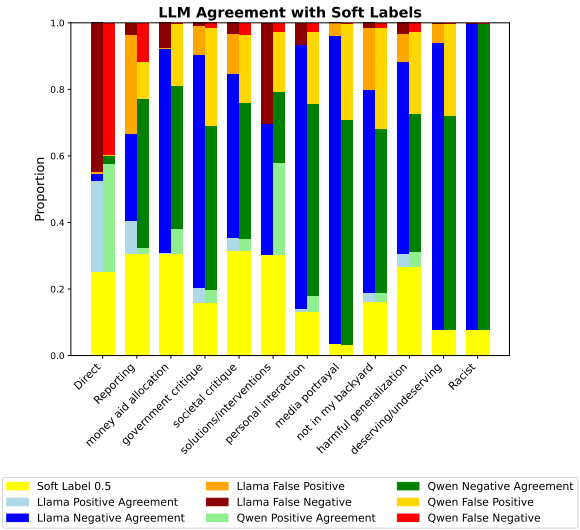

Figure 3: LLM Agreement with Human Annotators

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 losing 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. However, mitigation does a good job at reducing rhetorical questions; the kappa decreased by 0.17.

4.5 Example

Here is an example Anonymized Reddit Post: 'Most comments are saying how great it is to homeless (and it usually is) but are ignoring or unaware of the type* of homeless they plan to [STREET] here. Drug addicts and people with mental issues. If it were more homes for homeless and/or low income families, I wouldn't think twice about it but I'm very concerned about a facility housing drug addicts and people with mental issues just a couple hundred feet from a school in the middle of a residential neighborhood.'

The annotations are as follows:

Human Annotators: provide an observation, express their opinion, express others' opinions, societal critique (1 of 2 annotators), not in my backyard, harmful generalization, and deserving/undeserving.

Llama Classification: express their opinion, societal critique, and not in my backyard.

Qwen Classification: provide an observation, express their opinion, and harmful generalization.

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.

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, co-founder of a local non-profit, 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

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.

Alan Chan, Chinasa T Okolo, Zachary Terner, and Angelina Wang. 2021. The limits of global inclusion in ai development. *arXiv.org*.

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

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.

Alan Mislove, Sune Lehmann, Yong-Yeol Ahn, Jukka-Pekka Onnela, and J.Niels Rosenquist. 2011. Understanding the demographics of twitter users. *Proceedings of the Fifth International AAAI Conference on WEblogs and Social Media*.

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

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.

European Union. 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*.

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.

A Appendix

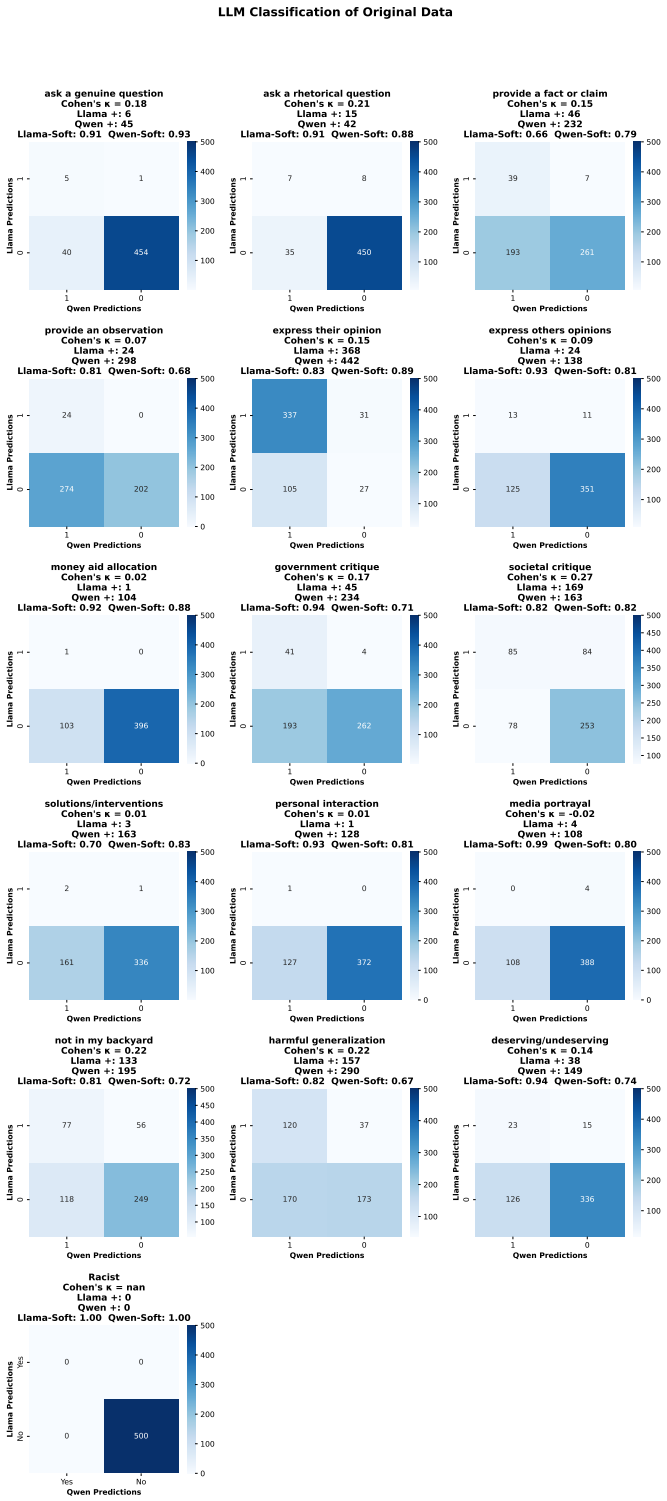

Figure 4: LLM Classification Confusion Matrices

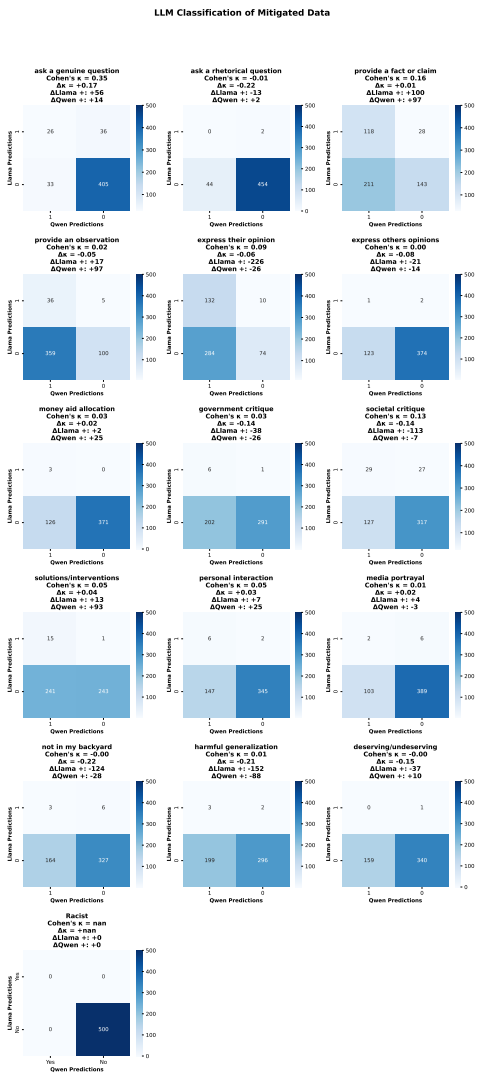

Figure 5: LLM Mitigated Confusion Matrices

Field	Soft 0.5	Soft 1	Soft 0	Llama+	Llama–	Qwen+	Qwen–	Mit. Llama+	Mit. Llama–	Mit. Qwen+	Mit. Qwen–
ask a genuine question	47	48	405	6	494	45	455	62	438	59	441
ask a rhetorical question	69	37	394	15	485	42	458	2	498	44	456
provide a fact or claim	181	196	123	46	454	232	268	146	354	329	171
provide an observation	138	98	264	24	476	298	202	41	459	395	105
express their opinion	103	350	47	368	132	442	58	142	358	416	84
express others opinions	123	27	350	24	476	138	362	3	497	124	376
money aid allocation	153	39	308	1	499	104	396	3	497	129	371
government critique	79	28	393	45	455	234	266	7	493	208	292
societal critique	157	37	306	169	331	163	337	56	444	156	344
solutions/interventions	151	152	197	3	497	163	337	16	484	256	244
personal interaction	66	37	397	1	499	128	372	8	492	153	347
media portrayal	17	1	482	4	496	108	392	8	492	105	395
not in my backyard	81	21	398	133	367	195	305	9	491	167	333
harmful generalization	134	36	330	157	343	290	210	5	495	202	298
deserving/undeserving	39	1	460	38	462	149	351	1	499	159	341
Racist	39	1	460	0	500	0	500	0	0	0	0

Table 2: Classification Counts by Field and Model

Soft Labels reflect partial (0.5), full (1), or no (0) human agreement. Llama/Qwen columns show model polarity (positive/negative), including mitigated variants.