Investigating Annotator Bias in Large Language Models for Hate Speech Detection

Amit Das[1], Zheng Zhang[2], Najib Hasan[3], Souvika Sarkar[3], Fatemeh Jamshidi[4], Tathagata Bhattacharya[5], Mostafa Rahgouy[6], Nilanjana Raychawdhary[6], Dongji Feng[7], Vinija Jain[8,9]*, Aman Chadha[8,9],*Mary Sandage[6], Lauramarie Pope[6], Gerry Dozier[6], Cheryl Seals[6]

[1]University of North Alabama, [2]Murray State University, [3]Wichita State University, [4]California State Polytechnic University Pomona, [5]Auburn University at Montgomery, [6]Auburn University, [7]Gustavus Adolphus College, [8]Stanford University, [9]Amazon GenAI
 Corresponding author: Amit Das (adas@una.edu)

Abstract

Data annotation, the practice of assigning descriptive labels to raw data, is pivotal in optimizing the performance of machine learning models. However, it is a resource-intensive process susceptible to biases introduced by annotators. The emergence of sophisticated Large Language Models (LLMs) presents a unique opportunity to modernize and streamline this complex procedure. While existing research extensively evaluates the efficacy of LLMs, as annotators, this paper delves into the biases present in LLMs when annotating hate speech data. Our research contributes to understanding biases in four key categories: gender, race, religion, and disability with four LLMs: GPT-3.5, GPT-4o, Llama-3.1 and Gemma-2. Specifically targeting highly vulnerable groups within these categories, we analyze annotator biases. Furthermore, we conduct a comprehensive examination of potential factors contributing to these biases by scrutinizing the annotated data. We introduce our custom hate speech detection dataset, *HateBiasNet*, to conduct this research. Additionally, we perform the same experiments on the ETHOS Mollas et al. (2022) dataset also for comparative analysis. This paper serves as a crucial resource, guiding researchers and practitioners in harnessing the potential of LLMs for data annotation, thereby fostering advancements in this critical field. The *HateBiasNet* dataset is available here: `https://github.com/AmitDasRup123/HateBiasNet`

Content Warning: This article features hate speech examples that may be disturbing to some readers.

1 Introduction

The growing widespread presence of online hate speech presents a critical challenge for maintaining safe and inclusive digital environments. Automated hate speech detection systems, powered by machine learning and large language models (LLMs), have emerged as vital tools to address this issue (Tan et al., 2024). These systems rely heavily on annotated datasets to train and evaluate their performance. However, the process of annotating hate speech is inherently subjective and influenced by the annotators' sociocultural, and personal biases. As a result, these biases can inadvertently be embedded into the datasets and, subsequently, into the detection models, raising concerns about fairness, accuracy, and generalizability. Despite the advancements in LLMs and their capabilities, annotator bias remains an underexplored yet significant factor affecting their performance and reliability in hate speech detection tasks.

LLMs offer a promising pathway toward transforming data annotation practices. Their ability to automate annotation tasks, ensure consistency across vast datasets, and adapt through fine-tuning or prompts tailored to specific domains significantly alleviates challenges inherent in traditional

---

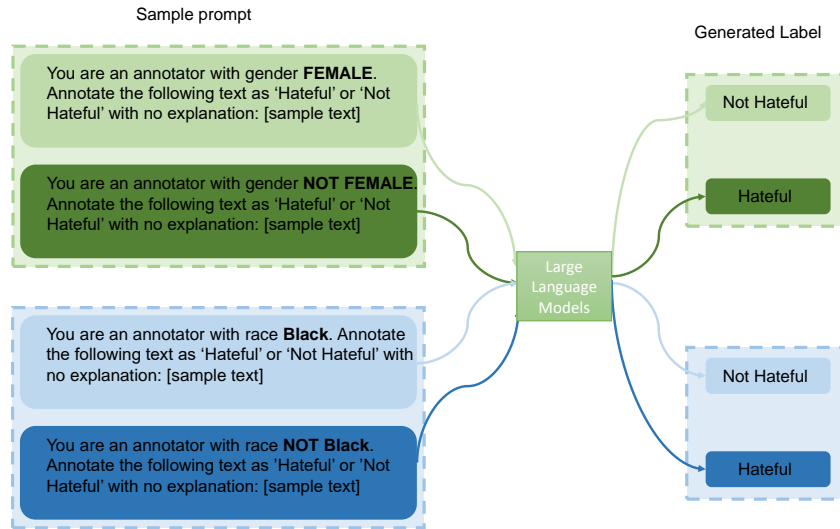*Work does not relate to position at Amazon.

Figure 1: Workflow diagram of our study, illustrating how varying biases can lead to different outcomes when annotating a sample text as hateful. We investigate annotator biases across four categories for hate speech detection using the following LLMs: GPT-3.5, GPT-4o, Llama-3.1, and Gemma-2.

annotation methodologies, thereby establishing a new standard for achievable outcomes in the realm of NLP.

However, data annotation with humans comes with the risk of annotator biases, both conscious and unconscious, that can significantly impact the downstream applications of AI systems. In this paper, we primarily focus on biases in LLMs for hate speech data annotation. Our paper explores four categories: gender, racial, religious, and disability-based bias. Specifically, we select the target groups that are highly vulnerable within the mentioned four categories and explore the annotator biases. Additionally, we provide a detailed analysis of the possible reasons for these biases by exploring the data being annotated. We compare the results across four LLMs: GPT-3.5, GPT-4o, Llama-3.1, and Gemm-2. We also explore why certain biases achieve higher accuracy compared to others.

Understanding the implications of these biases is crucial, as they can perpetuate harmful stereotypes and reinforce discrimination in automated systems. By systematically analyzing the biases present in LLMs, we aim to illuminate the mechanisms through which these biases emerge and manifest in the annotations they produce. This investigation not only highlights the ethical considerations involved in utilizing LLMs for sensitive tasks like hate speech detection but also provides insights for improving model training and annotation strategies to foster more equitable AI outcomes.

Moreover, the stakes are particularly high in the realm of hate speech detection, where the potential for mislabeling or biased labeling can lead to severe consequences, including the marginalization of vulnerable communities and the unjust suppression of free speech. As LLMs become increasingly integrated into content moderation systems, it is essential to ensure that the annotations they produce are not only accurate but also reflective of a fair and balanced perspective. By addressing these biases, we can pave the way for the development of more robust, transparent, and socially responsible AI systems, ultimately contributing to a safer online environment for all users. Serving as a critical guide, this paper aims to steer researchers toward exploring the potential of LLMs for data annotation, thereby facilitating future advancements in this essential domain.

Through rigorous data annotation, prompt engineering, quantitative and qualitative analysis, we aim to answer the following research questions:

RQ1: Does annotator bias exist in Large Language Models for hate speech detection?
RQ2: If it exists, what potential factors contribute to its existence?

RQ3: How can this problem be effectively mitigated?

To this end, our work makes the following contributions:

> **OUR CONTRIBUTIONS**
>
> ➠ Our research demonstrates that annotator bias is present in LLMs used for hate speech detection. This bias arises from the subjective interpretations of annotators, which influence the training data and consequently affect the model's performance. We provide empirical evidence illustrating how such biases skew detection results, leading to potential inaccuracies and unfair outcomes.
>
> ➠ In our research, we specifically examine four types of biases: gender, race, disability, and religion. Gender bias refers to the prejudiced treatment based on an individual's sex or gender identity. Race bias involves discriminary actions or attitudes towards individuals based on their racial or ethnic background. Disability bias encompasses unfair treatment of people with physical or mental impairments. Religion bias involves prejudices and discriminatory behaviors directed at individuals based on their religious beliefs or practices. Our study aims to analyze the prevalence and impact of these biases in various contexts.
>
> ➠ We delve into the underlying factors contributing to bias and propose a potential solution to address this issue. We analyze various aspects to uncover the root causes of bias and present a strategy aimed at mitigating its effects. Through our investigation, we aim to provide valuable insights into understanding and combatting bias in our study.

2 RELATED WORK

The advent of LLMs has revolutionized NLP tasks by enabling the development of more sophisticated and context-aware language understanding systems. Models such as BERT (Devlin et al., 2018), and their variants have demonstrated remarkable performance across a wide range of NLP tasks, including text classification, language generation, and question answering. These models leverage pre-training on large corpora followed by fine-tuning on task-specific data, allowing them to capture intricate linguistic patterns and semantic relationships.

Recent research has explored the use of LLMs for data annotation tasks, leveraging their ability to comprehend and generate human-like text. For instance, (Gururangan et al., 2020) proposed a framework for generating natural language explanations for machine learning models, facilitating the annotation of model predictions with interpretable justifications. Similarly, (Raffel et al., 2020) introduced a method for efficiently annotating speech data using GPT-2, demonstrating significant reductions in annotation time compared to traditional manual labeling approaches.

The increasing interest in leveraging Large Language Models as versatile annotators for various natural language tasks has been highlighted in recent research (Kuzman et al., 2023; Zhu et al., 2023; Ziems et al., 2024). (Wang et al., 2021) demonstrated that GPT-3 can significantly decrease labeling costs by up to 96% for both classification and generation tasks. Similarly, (Ding et al., 2023) conducted an assessment of GPT-3's effectiveness in labeling and data augmentation across classification and token-level tasks. Furthermore, empirical evidence suggests that LLMs can surpass crowdsourced annotators in certain classification tasks (Gilardi et al., 2023; He et al., 2023).

The investigation of social biases within Natural Language Processing (NLP) models constitutes a significant area of research. Previous studies have delineated two primary categories of biases and harms: allocational harms and representational harms (Blodgett et al., 2020; Crawford, 2017). Scholars have explored various methodologies to assess and alleviate these biases in both Natural Language Understanding (NLU) (Bolukbasi et al., 2016; Dixon et al., 2018; Zhao et al., 2018; Bordia & Bowman, 2019; Dev et al., 2021; Sun & Peng, 2021) and Natural Language Generation (NLG) tasks (Sheng et al., 2019; Dinan et al., 2019).

Within this body of literature, (Sun & Peng, 2021) proposed utilizing the Odds Ratio (OR) (Szumilas, 2010) as a metric to quantify gender biases, particularly in items exhibiting significant frequency disparities or high salience among genders. (Sheng et al., 2019) assessed biases in NLG model outputs conditioned on specific contextual cues, while (Dhamala et al., 2021) extended this analysis by incorporating real-world prompts extracted from Wikipedia. Several strategies (Sheng et al., 2020; Liu et al., 2021; Cao et al., 2022; Gupta et al., 2022) have been proposed to mitigate biases in NLG models, yet their applicability to closed API-based LLMs, such as ChatGPT, remains uncertain.

3 METHODOLOGIES

3.1 DATA COLLECTION AND ANNOTATION

The study initiates with the utilization of a hate speech lexicon sourced from Hatebase.org[1], comprising terms and expressions identified by online users as indicative of hate speech. Leveraging the Twitter API, we conducted a search for tweets containing lexicon terms, resulting in a corpus of 3003 tweets. Subsequently, three speech-language pathology graduate students were engaged for the purpose of data annotation. These annotators were tasked with categorizing each tweet into one of two classifications: hateful or not hateful. We name this dataset as ***HateBiasNet***.

Acknowledging the inherent vagueness in prior methodologies for annotating hate speech, as noted by (Schmidt & Wiegand, 2017), which often led to low agreement scores, our study took measures to enhance the clarity and consistency of the an- notation process. To achieve this, all annotators collaboratively formulated and refined annotation guidelines to ensure a shared understanding of hate speech. An explicit definition, accompanied by a detailed explanation, was provided to elucidate the concept further.

Annotators were instructed to consider not only the isolated words within a tweet but also the broader contextual usage of these terms. Emphasis was placed on discerning the intent behind the lan- guage and recognizing that the mere presence of offensive vocabulary did not inherently classify a tweet as hate speech. Each tweet underwent coding by three independent annotators, and the majority decision among them was employed to assign the final label. The annotation details are provided in the appendix.

3.2 DATA ANNOTATION BY LLMS

We then had our data annotated by the four LLMs. For the annotation, we first provided the annotator details, using direct prompt provided by (Das et al., 2024) for the annotation. One such prompt with the annotator being 'Female' is as follows. Note that `[Text]` refers to the input text to be annotated.

> You are an annotator with gender FEMALE. Annotate the following text as 'Hateful' or 'Not Hateful' with no explanation: [Text]

Category	Annotator Bias
Gender	Female vs. Not Female
Race	Asian vs. Not Asian
Race	Black vs. Not Black
Religion	Muslim vs. Not Muslim
Disability	Mental Disability vs. No Disability
Disability	Physical Disability vs. No Disability

Table 1: Annotator Biases in LLMs explored in this paper. With expert opinions, we selected six groups from four categories that face the most hateful comments on social media. We then explore the annotator bias in LLM annotation assuming the one annotator to be from one of the six categories, and one annotator not from that category.

3.3 ANNOTATOR BIASES

We used only the highly vulnerable groups on social media and used them as annotators. With expert opinions, we selected six groups from four categories that face the most hateful comments on social media. We then explore the annotator bias in LLM annotation assuming the one annotator to be from one of the six categories, and one annotator not from that category. Figure 1 depicts the workflow diagram of our work. The annotator biases we explored are given in Table 1.

---

[1]https://hatebase.org/

4 Results & Discussion

Along with **HateBiasNet**, we explored the same annotator bias on ETHOS (Mollas et al., 2022) dataset. We re-annotated the whole dataset using the four LLMs with the same experimental setup we used for annotating **HateBiasNet**. The analysis of data annotations by the LLMs revealed notable biases on both the datasets across each category. We observed a significant skew in the distribution of annotations towards the categories we used. Table 2 shows the mismatches between different annotator biases both on **HateBiasNet** and ETHOS dataset while annotating them with four LLMs: GPT-3.5, GPT-4o, Llama-3.1 and Gemma-2. It is observed that there are significant mismatches in both **HateBiasNet** and the ETHOS dataset. These findings underscore the presence of subjectivity and ambiguity in the LLM-based annotation process, highlighting the need for standardized guidelines and rigorous quality control measures.

Annotator Bias	Mismatch (%) for GPT-3.5 annotation		Mismatch (%) for GPT-4o annotation		Mismatch (%) for Llama-3.1 annotation		Mismatch (%) for Gemma-2 annotation	
	HateBiasNet Dataset	Ethos Dataset	*HateBiasNet* Dataset	Ethos Dataset	*HateBiasNet* Dataset	Ethos Dataset	*HateBiasNet* Dataset	Ethos Dataset
Asian vs. Not Asian	8.15	6.51	4.52	2.20	7.26	3.31	9.02	9.12
Black vs. Not Black	6.85	5.71	6.16	2.60	7.63	3.31	8.36	7.82
Female vs. Not Female	5.86	5.01	4.76	2.00	5.96	3.11	5.16	3.11
Mental Disability vs. No Disability	6.06	4.30	4.06	1.70	4.96	2.91	12.39	6.21
Physical Disability vs. No Disability	4.19	3.60	5.86	2.80	5.99	2.81	6.39	3.41
Muslim vs. Not Muslim	8.02	6.31	8.52	3.41	9.02	5.41	10.56	5.71

Table 2: Mismatches between different annotations when annotated by LLMs. It can be seen that for the ETHOS dataset, the biases are significantly reduced for GPT-4o annotation when compared to GPT-3.5, Llama-3.1 and Gemma-2 annotation.

4.1 GPT-3.5

In the context of GPT-3.5, annotator bias arises due to its pretraining on large, publicly available datasets, which embed societal biases, particularly in sensitive areas like hate speech. Since GPT-3.5 lacks alignment optimizations present in later models, it struggles with the nuanced classification of hate speech, leading to inconsistent annotations based on cultural and contextual variations.

An analysis of GPT-3.5 reveals significant differences in how annotators classify offensive language, influenced by their personal and cultural backgrounds. For example, Annotator 1 (Black) displayed leniency towards culturally contextual or ironic language, labeling statements like "The superior race but cannot stand in the sun without literally burning" as 'Not Hateful,' while Annotator 2 (Not Black) labeled the same statements as 'Hateful'. These discrepancies also extended to gender, sexuality, and disability-related content, where Annotator 1 (Female, Muslim, disabled) and Annotator 2 (Not Female, Not Muslim, no disability) showed different sensitivities. Annotator 1 was more tolerant of remarks related to gender and sexuality but more sensitive to offensive terms related to disability, whereas Annotator 2 exhibited opposite tendencies.

These findings highlight how annotators' personal identities, including race, religion, gender, and disability status, influence their interpretation of hate speech, leading to inconsistent classifications in GPT-3.5's output.

4.2 GPT-4o:

In analyzing GPT-4o's performance in hate speech detection, notable biases emerge despite improvements in its alignment and fine-tuning. These biases are linked to the model's broad training data and an overcompensation for ethical considerations, leading to inconsistencies in identifying offensive content, particularly in borderline or culturally specific cases.

The study reveals demographic and cultural disparities in hate speech annotation. For example, Asians consistently label derogatory terms such as 'cripple' or 'retard' as hateful due to their cultural and historical context, while non-Asians show more variability in their classifications. Significant differences also arise in how statements about black individuals, nationalities, disabilities, and religions are perceived. For instance, a Black annotator might classify a statement like "black Americans are more in danger than homosexuals" as hateful, while another annotator may not.

Gender and religious biases are also evident. Female annotators tend to classify texts involving gender, religion, or cultural insensitivity as hateful more frequently than male annotators. Similarly, a Muslim annotator may be more likely to flag content critical of religion as hateful compared to a non-Muslim annotator. Annotators with disabilities also show heightened sensitivity to disability-related comments, further underscoring the influence of personal identity and background on hate speech detection.

This analysis highlights the complex interplay of cultural, gender, and personal identity factors in shaping perceptions of hate speech in GPT-4o's annotations, revealing the challenges of achieving unbiased moderation across diverse demographic groups.

4.3 LLAMA-3.1:

In the case of Llama-3.1, a smaller research-focused model, bias may arise from its limited exposure to diverse and curated datasets, leading to skewed annotations that reflect dominant viewpoints in its training data. Compared to larger models like GPT-4, Llama-3.1 lacks robust alignment and fine-tuning, contributing to inconsistent or biased annotations, especially when dealing with nuanced hate speech.

Analysis of Llama-3.1 reveals significant annotator bias across demographic groups in hate speech detection. For example, a 'Black' annotator found the statement "hahahahahahahahaha europ fuck you fucking nazis" to be hateful due to its aggressive language toward Europeans, while a 'Not Black' annotator did not, possibly interpreting it as context-specific. Similarly, a 'Muslim' annotator flagged a statement with racial and inflammatory content as hateful, but a 'Not Muslim' annotator did not. Bias was also observed in relation to physical disability, where the 'Physical Disability' annotator perceived mocking language as hateful, but the 'No Disability' annotator did not. These cases highlight how personal identity shapes perceptions of hate speech, emphasizing the complexity of annotator bias in sensitive topics.

Annotator Bias	Accuracy (%) for GPT-3.5 annotation		Accuracy (%) for GPT-4o annotation		Accuracy (%) for Llama-3.1 annotation		Accuracy (%) for Gemma-2 annotation	
	HateBiasNet Dataset	Ethos Dataset	*HateBiasNet* Dataset	Ethos Dataset	*HateBiasNet* Dataset	Ethos Dataset	*HateBiasNet* Dataset	Ethos Dataset
Asian	74.09	80.56	71.49	86.87	68.6	**86.37**	72.53	77.86
Black	72.82	**80.96**	71.82	**87.47**	65.47	85.67	72.76	75.85
Female	75.65	79.75	75.95	85.27	78.42	83.67	70.3	**79.36**
Mental Disability	**76.49**	74.54	**76.22**	85.67	**80.62**	83.07	72.46	78.16
Muslim	73.85	79.65	69.49	87.07	55.21	85.97	72.19	78.95
Physical Disability	74.59	71.74	76.02	85.17	79.45	84.37	72.59	75.95
Not Asian	74.49	75.65	71.62	86.67	70.53	85.87	**73.43**	74.35
Not Black	73.42	77.85	69.96	87.27	68.45	86.17	73.05	73.25
Not Female	76.00	76.35	73.80	85.67	76.06	84.37	71.06	78.86
No Disability	75.57	72.54	76.02	86.17	79.99	83.17	71.86	76.35
Not Muslim	75.52	74.74	67.76	86.47	54.84	84.96	70.16	78.45

Table 3: Accuracy of different biases when compared to human annotation. It can be seen that for all the LLMs, the overall accuracies are higher for ETHOS dataset compared to *HateBiasNet*.

4.4 GEMMA-2:

In Gemma-2, bias can arise from the model's specialized focus and its reliance on narrow training data that may not capture the full spectrum of hate speech across diverse cultural and regional contexts. This limitation can result in the model disproportionately reflecting the perspectives of its training sources, leading to biased annotations. Additionally, dependence on specific hate speech detection guidelines or datasets may further perpetuate existing biases, resulting in inconsistent performance across different hate speech categories.

The model's bias is evident in how annotators from different demographic backgrounds perceive hate speech differently. For instance, a 'Black' annotator identifies racial undertones in a statement targeting Norwegians, viewing it as hateful, while a 'Not Black' annotator does not. Similarly, a 'Physical Disability' annotator finds a derogatory term offensive in a statement, whereas a 'No Disability' annotator is unaware of its negative connotations. Differences in perception also arise with mental disability-related comments, where a 'Mental Disability' annotator finds a stereotype about intelligence offensive, but a 'No Disability' annotator does not. These examples highlight how annotators' identities influence their interpretation of hate speech, revealing inherent biases in the model's output.
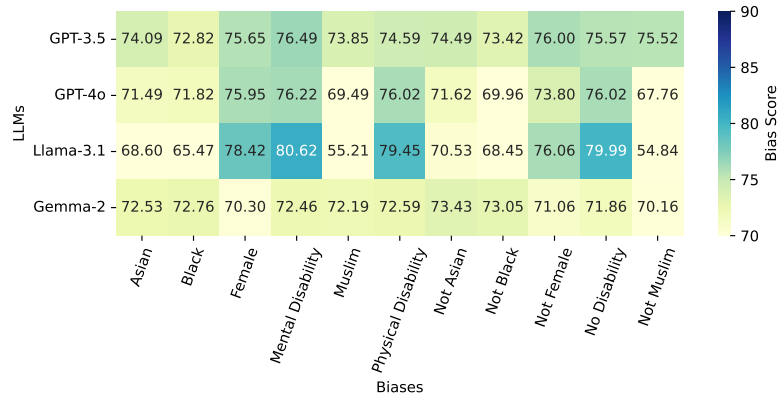
Figure 2: Heatmap of the ***HateBiasNet*** dataset illustrating the accuracy of 11 biases across 4 LLMs. Notably, GPT-3.5, GPT-4o, and Llama-3.1 demonstrate the highest accuracy for the 'Mental disability' bias. The word cloud of the dataset (Figure 4) suggests that specific keywords may influence annotation outcomes for these LLMs. Additionally, Llama-3.1 shows the highest accuracy overall for the 'Mental disability' bias among the 4 models.
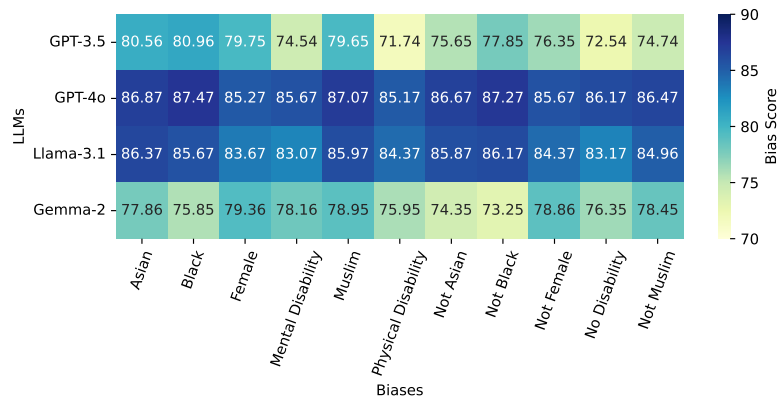

Figure 3: Heatmap of the ETHOS dataset depicting the accuracy of 11 biases across 4 LLMs. The bias 'Black' achieved the highest accuracy for both GPT-3.5 and GPT-4o, while 'Asian' exhibited the highest accuracy for Llama-3.1. The word cloud of the dataset (Figure 4) suggests that specific keywords may influence annotation results for these LLMs. Notably, GPT-4o and Llama-3.1 consistently outperformed GPT-3.5 and Gemma-2 across all biases. Among the LLMs, GPT-4o's performance on the 'Black' bias stands out as the highest overall accuracy.
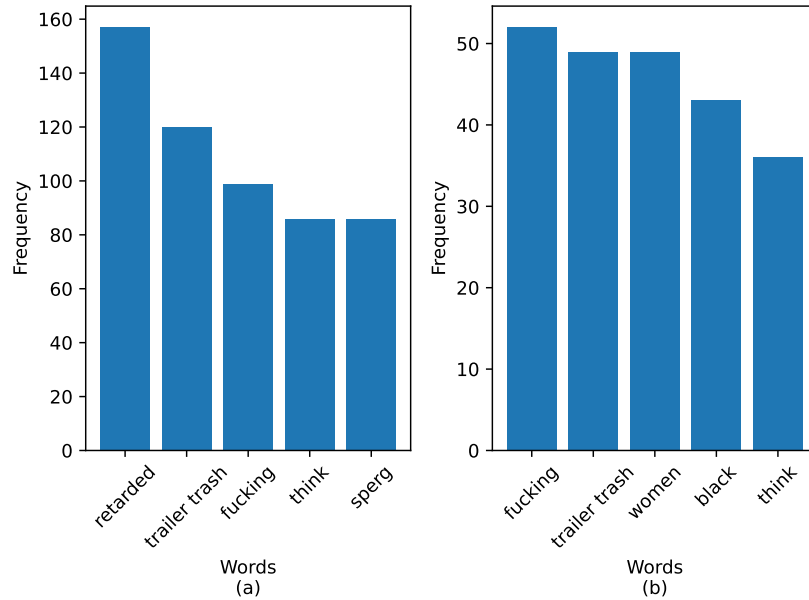
Figure 4: Word histogram (considering only the top 5 words) of (a) *HateBiasNet* and (b) ETHOS after removing the stopwords.

We proceeded to evaluate which bias yields the highest accuracy for data annotation by comparing the annotation results with the original human annotations. The results shown in Table 3 indicate that the Mental Disability bias achieves the highest accuracy on *HateBiasNet*. Additionally, Figure 4 illustrates the word histogram of *HateBiasNet* after the removal of stopwords. Notably, the most frequently occurring words are 'retarded' and 'trailer trash', both of which are closely associated with the mental disability bias.

For the ETHOS dataset, it can be seen that using Black bias gives the best result. Figure 4 shows the word histogram of Ethos dataset after removing the stopwords. It can be seen there is a clear relation between the keywords present in the dataset and the accuracy of data annotation by a particular group. We believe, although annotator bias exists in LLMs for hate speech detection, but selecting the correct prompt for the annotation can help mitigate this problem. Figures 2 and 3 show the heatmaps of the 11 biases across the the 4 LLMs for the *HateBiasNet* and the Ethos datasets respectively.

5   CONCLUSION

Our research highlights the presence of annotator biases in hate speech detection using both GPT-3.5 and GPT-4o, opening avenues for future investigation. One potential direction involves mitigating these biases by incorporating specific rules into the LLMs while training or prompting annotators to prevent biased outputs. Additionally, exploring broader aspects of the problem statement through enhanced language style or lexical content analyses holds promise.

The advent of LLMs like ChatGPT has introduced novel applications such as data annotation. However, our study underscores the risk of biases emerging when LLMs are directly utilized for annotation tasks. We meticulously assess four types of biases in LLM-assisted hate speech detection, revealing the propagation and amplification of harmful biases in annotations.

Our findings emphasize the need for cautious utilization of AI-assisted data annotation to counteract biases effectively. We advocate for the development of comprehensive policies governing the use of LLMs in real-world scenarios. Furthermore, we call for continued research into identifying and mitigating fairness issues in data annotation with LLMs, as understanding and addressing underlying biases are imperative for reducing potential harms in future LLM research endeavors.

REFERENCES

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. Language (technology) is power: A critical survey of "bias" in NLP. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5454–5476, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.485. URL https://aclanthology.org/2020.acl-main.485.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29, 2016.

Shikha Bordia and Samuel R Bowman. Identifying and reducing gender bias in word-level language models. *arXiv preprint arXiv:1904.03035*, 2019.

Yang Trista Cao, Yada Pruksachatkun, Kai-Wei Chang, Rahul Gupta, Varun Kumar, Jwala Dhamala, and Aram Galstyan. On the intrinsic and extrinsic fairness evaluation metrics for contextualized language representations. *arXiv preprint arXiv:2203.13928*, 2022.

Kate Crawford. The trouble with bias. In *Conference on Neural Information Processing Systems, invited speaker*, 2017.

Amit Das, Mostafa Rahgouy, Dongji Feng, Zheng Zhang, Tathagata Bhattacharya, Nilanjana Raychawdhary, Mary Sandage, Lauramarie Pope, Gerry Dozier, and Cheryl Seals. Offlandat: A community based implicit offensive language dataset generated by large language model through prompt engineering. *arXiv preprint arXiv:2403.02472*, 2024.

Sunipa Dev, Emily Sheng, Jieyu Zhao, Aubrie Amstutz, Jiao Sun, Yu Hou, Mattie Sanseverino, Jiin Kim, Akihiro Nishi, Nanyun Peng, et al. On measures of biases and harms in nlp. *arXiv preprint arXiv:2108.03362*, 2021.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Jwala Dhamala, Tony Sun, Varun Kumar, Satyapriya Krishna, Yada Pruksachatkun, Kai-Wei Chang, and Rahul Gupta. Bold: Dataset and metrics for measuring biases in open-ended language generation. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pp. 862–872, 2021.

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

Bosheng Ding, Chengwei Qin, Linlin Liu, Yew Ken Chia, Boyang Li, Shafiq Joty, and Lidong Bing. Is GPT-3 a good data annotator? In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11173–11195, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.626. URL https://aclanthology.org/2023.acl-long.626.

Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 67–73, 2018.

Soumitra Ghosh, Asif Ekbal, Pushpak Bhattacharyya, Tista Saha, Alka Kumar, and Shikha Srivastava. Sehc: A benchmark setup to identify online hate speech in english. *IEEE Transactions on Computational Social Systems*, 10(2):760–770, 2022.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. Chatgpt outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30):e2305016120, 2023.

Umang Gupta, Jwala Dhamala, Varun Kumar, Apurv Verma, Yada Pruksachatkun, Satyapriya Krishna, Rahul Gupta, Kai-Wei Chang, Greg Ver Steeg, and Aram Galstyan. Mitigating gender bias in distilled language models via counterfactual role reversal. *arXiv preprint arXiv:2203.12574*, 2022.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. Don't stop pretraining: Adapt language models to domains and tasks. *arXiv preprint arXiv:2004.10964*, 2020.

Xingwei He, Zhenghao Lin, Yeyun Gong, Alex Jin, Hang Zhang, Chen Lin, Jian Jiao, Siu Ming Yiu, Nan Duan, Weizhu Chen, et al. Annollm: Making large language models to be better crowdsourced annotators. *arXiv preprint arXiv:2303.16854*, 2023.

Taja Kuzman, Igor Mozetič, and Nikola Ljubešić. Chatgpt: Beginning of an end of manual linguistic data annotation? use case of automatic genre identification, 2023.

Alisa Liu, Maarten Sap, Ximing Lu, Swabha Swayamdipta, Chandra Bhagavatula, Noah A Smith, and Yejin Choi. Dexperts: Decoding-time controlled text generation with experts and anti-experts. *arXiv preprint arXiv:2105.03023*, 2021.

Ioannis Mollas, Zoe Chrysopoulou, Stamatis Karlos, and Grigorios Tsoumakas. Ethos: a multi-label hate speech detection dataset. *Complex & Intelligent Systems*, 8(6):4663–4678, 2022.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020.

Anna Schmidt and Michael Wiegand. A survey on hate speech detection using natural language processing. In *Proceedings of the fifth international workshop on natural language processing for social media*, pp. 1–10, 2017.

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

Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. Towards controllable biases in language generation. *arXiv preprint arXiv:2005.00268*, 2020.

Jiao Sun and Nanyun Peng. Men are elected, women are married: Events gender bias on wikipedia. *arXiv preprint arXiv:2106.01601*, 2021.

Magdalena Szumilas. Explaining odds ratios. *Journal of the Canadian academy of child and adolescent psychiatry*, 19(3):227, 2010.

Zhen Tan, Dawei Li, Song Wang, Alimohammad Beigi, Bohan Jiang, Amrita Bhattacharjee, Mansooreh Karami, Jundong Li, Lu Cheng, and Huan Liu. Large language models for data annotation: A survey. *arXiv preprint arXiv:2402.13446*, 2024.

Shuohang Wang, Yang Liu, Yichong Xu, Chenguang Zhu, and Michael Zeng. Want to reduce labeling cost? GPT-3 can help. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2021*, pp. 4195–4205, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-emnlp.354. URL https://aclanthology.org/2021.findings-emnlp.354.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Gender bias in coreference resolution: Evaluation and debiasing methods. *arXiv preprint arXiv:1804.06876*, 2018.

Yiming Zhu, Peixian Zhang, Ehsan-Ul Haq, Pan Hui, and Gareth Tyson. Can chatgpt reproduce human-generated labels? a study of social computing tasks. *arXiv preprint arXiv:2304.10145*, 2023.

Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. Can large language models transform computational social science? *Computational Linguistics*, pp. 1–55, 2024.

A APPENDIX

Here we explain the details of the ***HateBiasNet*** dataset creation and the results of the 4 LLMs.

A.1 ***HATEBIASNET*** DETAILS

Out of 3003 tweets, 2076 were annotated 'Hateful' and remaining 927 tweets were annotated 'Not Hateful' by the student annotators. The annotators achieved an average Cohen's Kappa score of 0.76. We paid each student above the national average per hour for doing the task.

A.2 STUDENT DATA ANNOTATION INSTRUCTIONS

1. You will be working with the 'Sample.csv' file.

2. The first column contains the tweets to be annotated. The second column is 'Label'. And the third column is for 'Category'. There is also an additional column, 'Comments', for comments.

3. You will not be changing anything in the first column. These are the data downloaded from X (Twitter).

4. After reading each tweet from the first column, you will be deciding whether a tweet is 'Hateful' or 'Not Hateful', and you will be putting it in the second column (titled 'Label') for the corresponding tweet. So basically, the second column, 'Label', will contain two options: 1. Hateful 2. Not Hateful. You can use the following definition of hate (Ghosh et al., 2022) to decide whether a tweet is hateful or not:

    Hate: A hate tweet contains a directed insult(s), vulgar language to denigrate a target or words that instigate or support violence. Furthermore, the simple use of offensive language such as slang and slurs on does not automatically result in a tweet of the type hate.

    Non-Hate: All other tweets that do not fall in the hate category are non-hate tweets.

5. If the 'Label' of a particular tweet is 'Not Hateful', leave the third column ('Category') blank.

6. To decide whether the text is hateful, check specifically two points: 1. The target of the hate and 2. The keywords used in the text.

7. There could be some non-English words in those cases to annotate by understanding the text's overall meaning.

8. The 'Sample Annotation.csv' file contains nearly 50 annotated tweets and the 'List of words.txt' file contains the list of words used to download the tweets. Please check these files before starting the annotation.

9. Lastly, you have an additional column, 'Comments' where you can add any comment if you want to. This section is totally optional.

A.3 KEYWORDS FOR DOWNLOADING TWEETS

In this section, we enlist the keywords used for downloading tweets to construct ***HateBiasNet***. These keywords were chosen to capture a wide range of offensive and derogatory language, enabling us to compile a comprehensive dataset for studying hate speech patterns on social media. Table 5 presents sample texts annotated by human annotators of various demographic groups and by two versions of a large language model (GPT-3.5 and GPT-4o). The table illustrates the variation and potential biases in annotations. Table 4 presents a selection of sample texts from the dataset accompanied by human-provided annotations.

htbp]

Text	Label (Annotation by Human)
God help us. The UK is completely retarded. Labour/Conservative just 2 sides of the same coin.	Hateful
This has to be the most retarded fucking post I've seen all year on X	Hateful
My man had Keith Bogans at the shooting guard. Can't believe he was able to win with that	Not Hateful
The sport of kings, funded by punters robbed by spivs.	Not Hateful
She's trailer trash. That mouth is like fingernails on a chalkboard.	Hateful
Rhyhiem. I saw Yardie stick a gun in somebody mouth and that took me out of it.	Not Hateful
Ooh as a yobbo I'm a meat pie and sauce kind of guy	Not Hateful

Table 4: Sample texts from ***HateBiasNet*** with human annotation.

> **Keywords:** bogan, bludger, hayseed/s, hoodrat/s, redneck/s, spiv/s, trailer trash, trailer park trash, yardie, yobbo/s, anchor baby, beaner/s, bull dyke/s, charva, chonkey, cripple/s, darky, eshays, gimp/s/y, goober, jigaboo, libtard, mongs, octaroon, peckerwood, quadroon, retard/s/ed, shemale, sperg, tard/s/ed, tranny, wetback, yokel, beaner, border hopper/s, border jumper/s, butt pirate/s, chinaman, ching chongs, coloreds, cunts, curry muncher, darkie/s, dindu nuffin, dune coon, dyke/s, fag, fagbag, fudgepacker, ghey/s, gypo, heebs, hilbilly, honkie/s, jiggaboo, jungle bunny, kikes, knacker, moon cricket/s, mud duck, mud shark, muzzie, ofay, papist/s, pickaninny, plastic paddy, pommie, tranny, whigger

A.4 ETHOS DATASET

The ETHOS dataset is designed for hate speech detection and is derived from YouTube and Reddit comments, which have been validated using a crowdsourcing platform. It comprises two subsets: one intended for binary classification and the other for multi-label classification. The binary classification subset includes 998 comments.

A.5 RESULTS OF THE BIASES ON THE ***HateBiasNet*** DATASET

In the ***HateBiasNet*** dataset, the 'Mental disability' bias recorded the highest accuracy among the biases evaluated across the LLMs GPT-3.5, GPT-4o, and Llama-3.1. Notably, Llama-3.1 exhibited the highest accuracy for the 'Mental disability' bias among all LLMs. The word cloud generated from the ***HateBiasNet*** dataset highlights "retarded" as the most frequently occurring word, which is directly associated with the 'Mental disability' bias. This suggests a potential correlation between accuracy and the presence of specific keywords in the dataset for these three LLMs. Figure 5 shows the line graph of the ***HateBiasNet*** dataset displaying 11 biases across the 4 LLMs.

A.6 RESULTS OF THE BIASES ON THE ETHOS DATASET

In the Ethos dataset, the bias 'Black' achieved the highest accuracy for both GPT-3.5 and GPT-4o, while the 'Asian' bias recorded the highest accuracy for Llama-3.1. Notably, GPT-4o demonstrated the highest overall accuracy for the 'Black' bias among all LLMs. Additionally, all LLMs showed higher average accuracy on the Ethos dataset compared to the ***HateBiasNet*** dataset. A word cloud generated from the Ethos dataset identified 'trailer trash' as the second most frequently occurring term (following the word 'fucking'), which is directly linked to the racial bias ('Black' and 'Asian'). This observation again suggests a potential correlation between accuracy and the presence of specific keywords in the dataset for these three LLMs. Figure 6 illustrates a line graph of the Ethos dataset, showcasing 11 biases across the four LLMs.
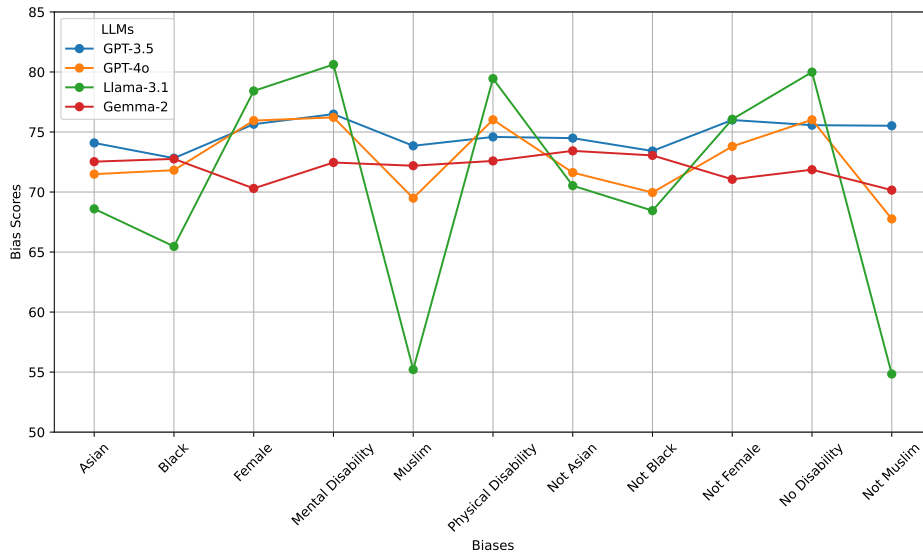
A.7 SAMPLE ANNOTATION RESULTS

Figure 5: Line graph of the ***HateBiasNet*** dataset displaying 11 biases across the 4 LLMs. Notably, for GPT-3.5, GPT-4o, and Llama-3.1, the 'mental disability' bias achieved the highest accuracy. The word cloud of the dataset (Figure 4) suggests that the presence of specific keywords may influence the annotation outcomes for these three LLMs.
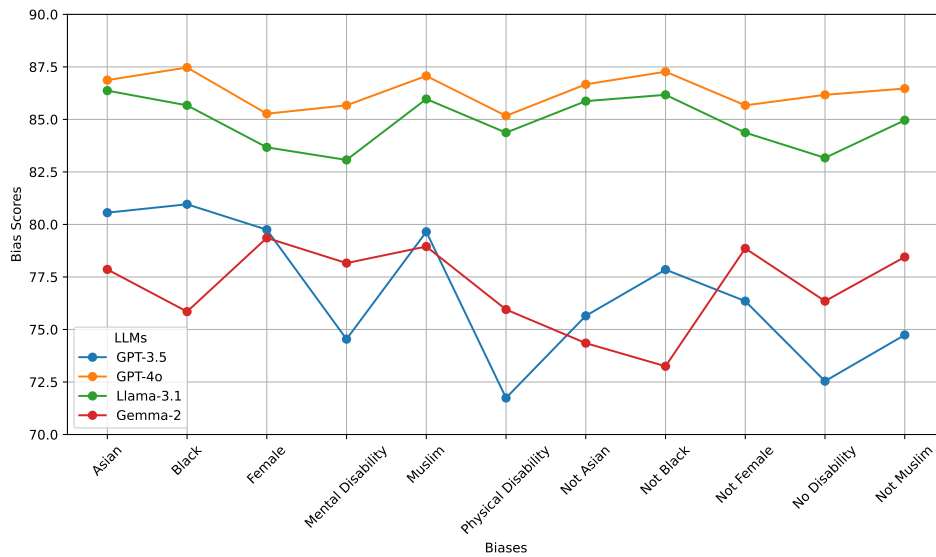

Figure 6: Line graph of the Ethos dataset showing the performance of 11 biases across 4 LLMs. It is evident that the bias 'Black' achieved the highest accuracy for GPT-3.5 and GPT-4o, while 'Asian' showed the highest accuracy for Llama-3.1. The word cloud of the dataset (see Figure 4) suggests that the presence of specific keywords may influence annotation outcomes for these LLMs. Additionally, it is noticeable that GPT-4o and Llama-3.1 consistently outperformed GPT-3.5 and Gemma-2 across all biases for the Ethos dataset.

Text	Dataset	Annotator	Annotation	Annotator	Annotation	LLM
Hello I'm calling from tech support *with Indian accent*	ETHOS	Asian	Hateful	Not Asian	Not Hateful	GPT-4o
having ADHD is so funny bc i think im a genius but i also know that im severely retarded	HateBiasNet	Asian	Hateful	Not Asian	Not Hateful	GPT-4o
Killing me to see gays treated like that. I think lesbians are being more privileged by people	ETHOS	Asian	Not Hateful	Not Asian	Hateful	GPT-3.5
why the hell is this Norwegian speaking about our LAWS? SPEAK FOR YOUR OWN Scandinavian scam	ETHOS	Black	Hateful	Not Black	Not Hateful	Gemma-2
hahahahahahahahaha europ fuck you fucking nazis	ETHOS	Black	Hateful	Not Black	Not Hateful	Llama-3.1
All women love listening to me sperg out, they just hate everything else about me	HateBiasNet	Black	Hateful	Not Black	Not Hateful	GPT-4o
Sweet and cute yet will cripple a beta with a snap of a finger.	HateBiasNet	Black	Not Hateful	Not Black	Hateful	GPT-3.5
Fucking Canada...stupid. I think the cold shrinks their brains.	ETHOS	Female	Hateful	Not Female	Not Hateful	GPT-4o
This sick, disgusting, white terrorist/fascist killed more than 50 innocent people.	ETHOS	Muslim	Hateful	Not Muslim	Not Hateful	Llama-3.1
I see the depression caused her to gain more weight.	ETHOS	Physical Disability	Hateful	No Disability	Not Hateful	GPT-4o
American people are dumbass. They don't even know geography	ETHOS	Mental Disability	Hateful	No Disability	Not Hateful	Gemma-2
What imaginary gender are you?	ETHOS	Muslim	Not Hateful	Not Muslim	Hateful	GPT-3.5
The Resistance avenge women and children killed by US and Allies !!!	ETHOS	Physical Disability	Not Hateful	No Disability	Hateful	GPT-3.5

Text	Dataset	Annotator	Annotation	Annotator	Annotation	LLM
You look like someone who would do an electric wheelchair race with Stephen Hawking.	ETHOS	Physical Disability	Not Hateful	No Disability	Hateful	GPT-3.5
Houston is an important petro chemical hub. Bomb Houston and you could cripple multiple industries.	HateBiasNet	Muslim	Hateful	Not Muslim	Not Hateful	Llama-3.1
Locked rubber gimp. Who said Locktober is going to end today?? I stay locked	HateBiasNet	Physical Disability	Hateful	No Disability	Not Hateful	Gemma-2
Honestly the way football is turning is just sad, people are more bothered about filming themselves than the moments, so called celebs on the pitch during titles are won, and now gimps like this at a so called prestigious event. Embarrassing	HateBiasNet	Mental Disability	Hateful	No Disability	Not Hateful	GPT-4o
Houston is an important petro chemical hub. Bomb Houston and you could cripple multiple industries.	HateBiasNet	Muslim	Not Hateful	No Muslim	Hateful	GPT-3.5
i was laughing at this boy on crutches ystd now im cripple & hopping roun dis bih?	HateBiasNet	Physical Disability	Hateful	No Disability	Not Hateful	Llama-3.1
No i'm dealing with a brainwashed yankee. A retard basically	HateBiasNet	Mental Disability	Not Hateful	No Disability	Hateful	Gemma-2
Houston is an important petro chemical hub. Bomb Houston and you could cripple multiple industries.	HateBiasNet	Muslim	Not Hateful	No Muslim	Hateful	GPT-3.5

Text	Dataset	Annotator	Annotation	Annotator	Annotation	LLM
It's scary how "Technical Debt Cripples Companies And Threatens To Stifle Innovation" - learn more in this segment on with vFunction CEO	HateBiasNet	Physical Disability	Hateful	No Disability	Not Hateful	GPT-3.5
You've wanted to visit for years gimp & shared your deepest needs online. It's time. Let's get started.	HateBiasNet	Mental Disability	Not Hateful	No Disability	Hateful	GPT-3.5

Table 5: Sample Texts with annotation biases while annotating by the LLMs.