

Investigating Annotator Bias in Large Language Models for Hate Speech Detection

Amit Das¹, Zheng Zhang¹, Fatemeh Jamshidi¹, Vinija Jain^{2,3*}, Aman Chadha^{2,3*},
Nilanjana Raychawdhary¹, Mary Sandage¹, Lauramarie Pope¹,
Gerry Dozier¹, Cheryl Seals¹

¹Auburn University, AL, USA

²Stanford University, USA, ³Amazon AI, USA

azd0123@auburn.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), like ChatGPT 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, specifically GPT 3.5 and GPT 4o when annotating hate speech data. Our research contributes to understanding biases in four key categories: gender, race, religion, and disability. 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, *HateSpeechCorpus*, 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 *HateSpeechCorpus* dataset is available here: <https://github.com/AmiDasRup123/HateSpeechCorpus>

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

1 Introduction

Within the multifaceted domain of machine learning and Natural Language Processing (NLP), data annotation emerges as a critical juncture, extending beyond simple label assignment to encompass a

wide array of supplementary predictive data. This intricate procedure entails a series of nuanced steps: initially, sorting raw data with class or task labels for primary classification, followed by the inclusion of intermediate labels to enrich contextual understanding; subsequently, assigning confidence scores to gauge the reliability of annotations; further, incorporating alignment or preference labels to tailor outputs to specific criteria or user needs; then, annotating entity relationships to illuminate interactions within datasets; delineating semantic roles to uncover the underlying functions of entities within sentences; and finally, tagging temporal sequences to capture the chronological flow of events or actions.

Data annotation poses notable challenges for contemporary machine learning models, owing to the complexity, subjectivity, and diversity of the data involved. It demands domain expertise and entails labor-intensive manual labeling of extensive datasets. Additionally, the process of annotating data is often labor-intensive, requiring significant human effort and expertise. In recent years, the emergence of LLMs, such as OpenAI's GPT series and Google's BERT, has revolutionized natural language processing tasks by demonstrating remarkable capabilities in understanding and generating human-like text.

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 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 pa-

*Work does not relate to position at Amazon.

per, 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 of these biases by exploring the data being annotated. Serving as a critical guide, this paper aims to steer researchers towards 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 discriminatory 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), GPT (Radford 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; Ziemis 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 crowd-sourced 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 and Bowman, 2019; Dev et al., 2021; Sun and Peng, 2021) and Natural Language Gener-

ation (NLG) tasks (Sheng et al., 2019; Dinan et al., 2019).

Within this body of literature, (Sun and 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.org¹, 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 *HateSpeechCorpus*.

Acknowledging the inherent vagueness in prior methodologies for annotating hate speech, as noted by (Schmidt and Wiegand, 2017), which often led to low agreement scores, our study took measures to enhance the clarity and consistency of the annotation 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 language and recognizing that the mere presence of offensive vocabulary did not inherently classify a tweet as hate speech. Each tweet underwent coding

¹<https://hatebase.org/>

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 ChatGPT

We then had our data annotated by ChatGPT. For the annotation, we first provided the annotator details, using direct prompt provided by (Das et al., 2024) for the annotation. For the annotation, we used the OpenAI API with both ‘gpt-3.5-turbo’ and ‘gpt-4o’. 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.

4 Results & Discussion

Along with *HateSpeechCorpus*, we explored the same annotator bias on ETHOS (Mollas et al.,

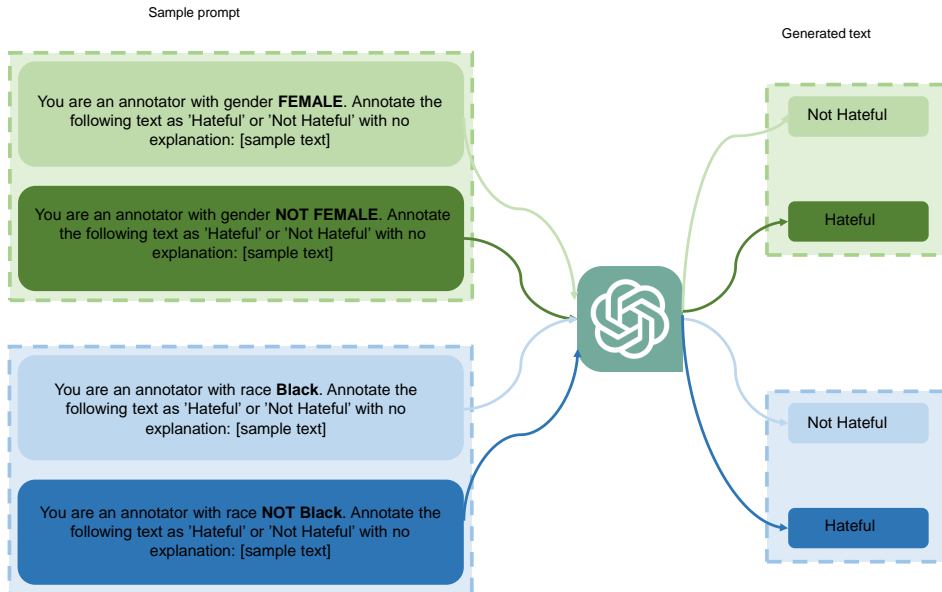


Figure 1: Workflow diagram of our work. It shows that different biases may produce different results while annotating a sample text as hateful. We explore annotator biases in four different categories for hate speech detection on both GPT-3.5 and GPT-4o.

2022) dataset. We re-annotated the whole dataset using both GPT-3.5 and GPT-4o with the same experimental setup we used for annotating *HateSpeechCorpus*. 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 *HateSpeechCorpus* and ETHOS dataset while annotating them with both GPT-3.5 and GPT-4o. It is observed that there are significant mismatches in both *HateSpeechCorpus* 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.

4.1 Racial Bias

4.1.1 Asian

In the analysis of hateful comments by GPT-4o, a significant disparity exists between Asians and non-Asians in perceiving and labeling offensive language. Asians often encounter remarks directly tied to ethnicity, such as comments on accents or appearances specific to their ethnic background. Conversely, non-Asians focus on broader social

issues or behaviors, with a differing set of terms and focal points in their hateful remarks. This divergence is attributed to the heightened sensitivity of Asians to certain derogatory terms due to their cultural and historical experiences, leading to more consistent labeling of terms like ‘cripple’ or ‘retard’ as hateful. Non-Asians, however, exhibit more variability in their classifications, likely due to different cultural understandings or personal thresholds for offensive content.

For GPT-3.5, distinct patterns emerge between ‘Asian’ and ‘Not Asian’ annotators in evaluating offensive language. The ‘Asian’ annotator tends to be more lenient, often focusing on explicit hatefulness and considering context heavily, thus less likely to label texts as ‘Hateful’ if mitigating context exists. In contrast, the ‘Not Asian’ annotator is more stringent, labeling potentially offensive terms as ‘Hateful’ regardless of context, indicating a higher sensitivity to broader social implications. This difference underscores the impact of cultural variations in interpreting sarcasm, humor, and social issues, leading to significant divergences in how annotators weigh context and language in the annotation process.

4.1.2 Black

In the analysis of labeling statements as ‘Hateful’ or ‘Not Hateful’, discrepancies between annotators,

particularly in GPT-4o, are apparent with issues concerning black individuals, specific nationalities, disabilities, and religions. For example, Annotator 2, who is 'Not Black' frequently labels statements addressing black individuals as 'Hateful', whereas similar comments about other groups might not receive the same label. A notable instance is the claim that "black Americans are more in danger than homosexuals", which Annotator 1, who is 'Black' considers 'Hateful', while Annotator 2 does not. Additionally, statements criticizing disabilities and religions also show varied labeling between the annotators. Annotator 1 tends to label derogatory terms such as 'gimp' and 'retard' as 'Hateful', reflecting a heightened sensitivity to offensive language, potentially due to personal or cultural awareness, whereas Annotator 2 often views such terms as colloquial and less offensive, suggesting a different cultural or racial background influencing their interpretations.

For GPT-3.5, Annotator 1 often does not mark statements containing cultural references, slurs, or potentially offensive language as 'Hateful', possibly due to a nuanced understanding of context and intent. Statements like 'The superior race' but cannot stand in the sun without literally burning the fuck up and "As a black person, I think that rappers should stop putting the n-word in the song. Plus, you should be proud of being White. I am proud of being black, so.." are marked as 'Not Hateful' by Annotator 1, recognizing potential irony, humor, or intra-community critique, while Annotator 2 labels them as 'Hateful'. Furthermore, terms like 'cripple', 'gimp', 'retarded', and 'sperg' are considered non-hateful by Annotator 1 in certain contexts, whereas Annotator 2 consistently finds them hateful. This pattern indicates that Annotator 1 has a higher threshold for what is deemed hateful, influenced by their cultural background and perception of offense, contrasting with Annotator 2's broader interpretation of offensive language, leading to more content being marked as 'Hateful' irrespective of context.

4.2 Gender

Research on GPT-4o reveals a potential gender-based difference in the perception of hateful speech, with female annotators showing a higher tendency to classify texts as 'hateful' when they involve aggressive, gender-related, religious, and cultural insensitivities. For instance, Annotator 1, who

is 'female', labels texts with aggressive language or general insults as hateful, such as "Fucking Canada...stupid. I think the cold shrinks their brains" and "It's not 'genderless' they just look like GIRLS", while Annotator 2, who is 'Not Female' does not. Annotator 1 also shows heightened sensitivity to gender and sexuality-related comments, like "Every time that blonde lady talks I lose twenty brain cells", labeling them as hateful more frequently than Annotator 2, who reserves the label for more explicit threats or overt hate speech. Additionally, Annotator 1 considers context heavily, sometimes marking casual or self-referential derogatory terms as not hateful, such as "not to brag but I'm pretty gay and retarded", indicating that intent and context are crucial in their assessments. Conversely, Annotator 2, who likely represents a non-female perspective, displays higher sensitivity to offensive language overall, suggesting a more conservative approach to annotation.

In contrast, research on GPT-3.5 shows that Annotator 2 ('Not Female') exhibits higher sensitivity to offensive language, particularly derogatory remarks concerning gender, sexuality, and other sensitive topics. Annotator 2 frequently labels content as hateful when it involves derogatory comments about women and LGBTQ+ individuals, such as "I didn't know future gay gym teachers had to take AIDS tests." On the other hand, Annotator 1 ('Female') tends to be more lenient, often classifying such content as 'Not Hateful', possibly perceiving it as humor or sarcasm. For example, Annotator 1 marked "Gas should be your preferred drink" as 'Not Hateful', while Annotator 2 deemed it 'Hateful'. This pattern extends to various terms and phrases, with Annotator 1 potentially overlooking offensive language if the overall message does not seem hateful, whereas Annotator 2 focuses on specific words and their potential to offend. This difference in sensitivity and interpretation suggests that Annotator 1's approach is influenced by a higher tolerance for offensive humor, while Annotator 2 applies a stricter criterion, considering the broader context and potential harm of the language used. This variance underscores how personal experiences and biases can shape the interpretation of potentially offensive content.

4.3 Religious Bias

The analysis of GPT-4o annotations reveals notable biases influenced by the religious backgrounds of

Annotator Bias	Mismatch (%) for GPT-3.5 annotation		Mismatch (%) for GPT-4 annotation	
	HateSpeech-Corpus	Ethos Dataset	HateSpeech-Corpus	Ethos Dataset
	Female vs. Not Female	5.86	5.01	4.76
Mental Disability vs. No Disability	6.06	4.30	4.06	1.70
Physical Disability vs. No Disability	4.19	3.60	5.86	2.80
Asian vs. Not Asian	8.15	6.51	4.52	2.20
Black vs. Not Black	6.85	5.71	6.16	2.60
Muslim vs. Not Muslim	8.02	6.31	8.52	3.41

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-4 annotation when compared to GPT-3.5 annotation.

two annotators, identified as a Muslim Annotator (annotator 1) and a Non-Muslim Annotator (annotator 2). annotator 1 tends to label content critical of religion, especially Christianity or Hinduism, as well as violent language, ethnic, and racial comments as hateful. For instance, annotator 1 finds the statement “Syria has a religion problem” hateful, whereas annotator 2 does not, and labels a comment about Korean appearance as hateful, which annotator 2 overlooks. Conversely, annotator 2 is less sensitive to these issues but more attentive to gender-related comments and specific political and disability-related content. For example, annotator 1 consistently marks derogatory terms such as ‘cripple’, ‘gimp’, ‘retard’, and ‘libtard’ as hateful due to their offensive nature, while annotator 2 tends to overlook them, suggesting a higher tolerance or different interpretation. These patterns indicate that annotator 1 has a heightened sensitivity to religious and ethnic content, often perceiving critiques or generalizations as more hateful, while annotator 2’s annotations reflect a belief that context can justify the use of certain aggressive or derogatory terms.

For GPT-3.5, the analysis of annotators’ responses to potentially hateful comments highlights significant divergences influenced by cultural and religious perspectives. annotator 1 shows higher tolerance for comments related to gender, sexuality, violence, general insults, and certain offensive terms like ‘cripple’, ‘gimpy’, ‘retard’, and ‘sperg’. For instance, comments such as “What imaginary gender are you?” and “Do the world a favor and pour some bleach into that glass” were labeled as ‘Not Hateful’ by annotator 1 but deemed ‘Hateful’ by annotator 2. annotator 2, identified as Not Muslim, consistently labels such comments as hateful, indicating greater sensitivity to issues concern-

ing gender, sexuality, violence, and degrading language. This suggests that annotator 1’s interpretations are influenced by their personal identification with specific religious and cultural contexts, leading to a more permissive stance on language that could be seen as offensive. In contrast, Annotator 2’s broader inclusivity and heightened sensitivity reflect a cultural perspective prioritizing the recognition and labeling of potentially harmful content. This disparity underscores how annotators’ backgrounds and cultural contexts significantly shape their perceptions of what constitutes hateful language.

4.4 Disability Based Bias

4.4.1 Physical Disability

In a study examining the labeling tendencies of annotators for GPT-4o and GPT-3.5 regarding hate speech, distinct differences were observed in their sensitivity to offensive language. Annotator 1 (‘annotator with Physical Disability’) from the GPT-4o study consistently marked comments related to disabilities, physical appearance, or general insults as hateful. This annotator displayed a heightened sensitivity to derogatory or offensive language, often labeling comments like “I see the depression caused her to gain more weight” and “Is this guy actually autistic or is he faking it” as hateful, even when Annotator 2 (‘annotator with No Disability’) did not. Annotator 1 also flagged terms such as ‘cripple’, ‘gimp’, ‘retarded’, and ‘sperg’ as hateful, reflecting a lower threshold for hate speech. In contrast, Annotator 2 was less sensitive to these terms, focusing more on gender issues, feminism, and explicit violence, which resulted in fewer instances being marked as hateful in the contexts emphasized by Annotator 1. This divergence underscores the influence of personal sensitivities and contextual

Annotator Bias	Accuracy (%) for GPT-3.5 annotation		Accuracy (%) for GPT-4o annotation	
	HateSpeech-Corpus	Ethos Dataset	HateSpeech-Corpus	Ethos Dataset
Asian	74.09	80.56	71.49	86.87
Black	72.82	80.96	71.82	87.47
Female	75.65	79.75	75.95	85.27
Mental Disability	76.49	74.54	76.22	85.67
Physical Disability	74.59	71.74	76.02	85.17
Not Asian	74.49	75.65	71.62	86.67
Not Black	73.42	77.85	69.96	87.27
Not Female	76.00	76.35	73.80	85.67
No Disability	75.57	72.54	76.02	86.17

Table 3: Accuracy of different biases when compared to human annotation. It can be seen that for the ETHOS dataset, the accuracies are significantly higher for GPT4o annotation when compared to GPT-3.5 annotation. For the HateSpeech-Corpus, the ‘mental disability’ bias achieved the highest accuracy whereas for the ETHOS dataset the ‘Black’ bias achieved the highest accuracy for both GPT-3.5 and GPT-4o.

interpretations on the assessment of hate speech.

Similarly, in the GPT-3.5 study, Annotator 1 exhibited a heightened sensitivity to aggressive language and terms related to physical or mental impairments. This annotator labeled comments such as “I hate being alone at night. It’s fucking creepy” and terms like ‘cripple’ and ‘retarded’ as hateful, regardless of context. In contrast, Annotator 2 showed a more context-dependent approach, often not marking such terms as hateful without additional context. For example, the term ‘cripples’ in a technical context was labeled hateful by Annotator 1 but not by Annotator 2. These findings indicate that annotators’ personal experiences and sensitivities significantly influence their judgments on hate speech, with Annotator 1 demonstrating greater sensitivity to potentially harmful language and Annotator 2 displaying a reliance on contextual cues to determine the hatefulness of comments.

4.4.2 Mental Disability

For GPT-4o, annotator 1 (‘annotator with Mental Disability’) shows a heightened sensitivity to language that implies mental health conditions, derogatory terms, offensive language, implicit bias, and aggressive contexts, often categorizing such phrases as ‘Hateful’. For instance, phrases like “Is this guy actually autistic or is he faking it” and “I see the depression caused her to gain more weight” are marked as hateful due to their references to mental health conditions. Violent and aggressive contexts, such as “HE SHOT THE KID!?!” and “The Resistance avenge women and children...” are also labeled as hateful. Annotator 1 also consid-

ers derogatory remarks and dismissive comments like “Wow feminists are real hypocrites...” and “What imaginary gender are you?” as hateful, recognizing the implicit harm in terms like ‘cripple’, ‘retard’, ‘mong’, and ‘sperg’. In contrast, Annotator 2 (‘annotator with No Disability’) demonstrates a broader tolerance, often labeling the same texts as ‘Not Hateful’, potentially overlooking the derogatory nature or historical baggage of such terms. Examples include “I’m a Cripple but I challenge #StarmerOut to see who has more integrity or even an hour in the ring” and “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”. This difference in sensitivity highlights the bias in their annotations, with Annotator 1’s heightened awareness and sensitivity towards derogatory language leading to more frequent ‘Hateful’ labels compared to Annotator 2.

For GPT-3.5, Annotator 1 consistently labels statements mentioning disabilities, such as “You look like someone who would do an electric wheelchair race with Stephen Hawking” and “I have a schizophrenia and I fight alone,” as ‘Hateful’. Conversely, Annotator 2 does not uniformly classify these statements as ‘Hateful’, suggesting that Annotator 1 perceives any mention of disability as offensive, whereas Annotator 2 might not view these in the same light. This pattern extends to other sensitive topics like race, gender, sexual orientation, and religion, with Annotator 1 demonstrating heightened sensitivity to potentially dis-

crimatory or critical comments, such as “Society is fucked for sure. And that’s everyone’s fault” and “Women cannot create children, only men can do that. Women can bear children”, labeling them ‘Hateful’ more frequently. Meanwhile, Annotator 2’s interpretations indicate a higher threshold for harmful language, possibly evaluating comments with more contextual leniency. Although Annotator 2 flags terms like ‘cripple’, ‘gimp’, and ‘retarded’ as ‘Hateful’ due to their derogatory nature, Annotator 1 may not view these terms as inherently offensive or harmful in context, resulting in more ‘Not Hateful’ classifications. This disparity underscores the subjective nature of labeling speech and the influence of personal biases, with Annotator 2 applying a more stringent standard to potentially derogatory language.

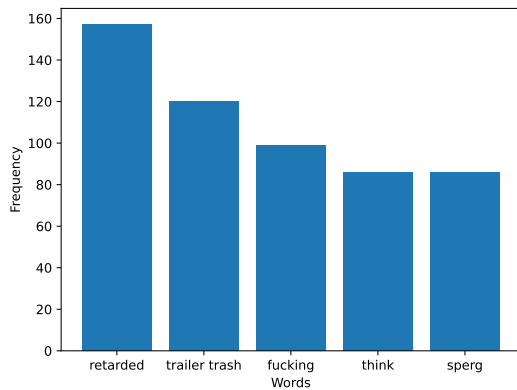


Figure 2: Word histogram (considering only the top 5 words) of *HateSpeechCorpus* after removing the stopwords.

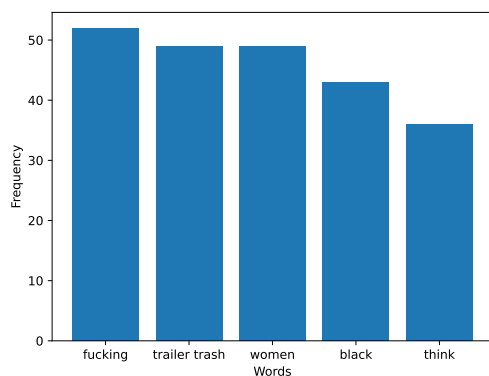


Figure 3: Word histogram (considering only the top 5 words) of the ETHOS dataset after removing the stopwords.

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 *HateSpeechCorpus*. Additionally, Figure 2 illustrates the word histogram of *HateSpeechCorpus* 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 3 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 ChatGPT for hate speech detection, but selecting the correct prompt for the annotation can help mitigate this problem.

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.

Limitations

While this study provides valuable insights into the biases present in Large Language Models like ChatGPT when annotating hate speech data, several limitations should be acknowledged. The annotated dataset used for analysis is static, and does not account for the evolving nature of hate speech, which may affect the relevance of the findings to future datasets or emerging forms of hate speech. Additionally, the study assumes a one-way interaction with the LLM acting solely as an annotator, without considering potential human-AI collaboration that could mitigate some biases. The research does not account for the diversity of human annotators involved in training the LLMs, whose inherent biases could influence the model's performance. Furthermore, the study may not fully capture the complexity of contextual nuances in hate speech, particularly subtle or coded forms that are difficult for LLMs to detect. While primarily employing quantitative methods, the study might overlook qualitative nuances better captured through in-depth qualitative analysis. Ethical considerations related to the deployment of LLMs for hate speech annotation, such as the potential harms of misclassification and societal impacts, are not extensively discussed. Finally, the study identifies biases without proposing or evaluating specific strategies for their mitigation, underscoring the need for future research to explore effective methods for reducing bias and enhancing the fairness and reliability of LLMs in hate speech detection.

Ethics Statement

This research investigates biases in LLMs, during hate speech annotation, focusing on gender, race, religion, and disability. We ensured data privacy by using publicly available data or data collected from social media and anonymizing identifiable information. Extensive measures were taken to minimize harm, including providing support resources for annotators handling sensitive content. We undertook extensive measures in annotating the *Hate-SpeechCorpus* dataset, securing IRB approval, and adhering to legal and ethical guidelines throughout the data handling and annotation process. Our methodology rigorously identified and explored biases. We maintained transparency and accountability by thoroughly documenting our processes, enabling reproducibility and critical evaluation. Emphasizing responsible use, we advocate for ethical

guidelines in deploying LLMs for data annotation. Continuous ethical review ensured alignment with ethical standards and societal values, aiming to inform and guide future efforts in creating fairer and more accurate hate speech detection systems.

References

- Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. [Language \(technology\) is power: A critical survey of “bias” in NLP](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.
- Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29.
- Shikha Bordia and Samuel R Bowman. 2019. Identifying and reducing gender bias in word-level language models. *arXiv preprint arXiv:1904.03035*.
- Yang Trista Cao, Yada Pruksachatkun, Kai-Wei Chang, Rahul Gupta, Varun Kumar, Jwala Dhamala, and Aram Galstyan. 2022. On the intrinsic and extrinsic fairness evaluation metrics for contextualized language representations. *arXiv preprint arXiv:2203.13928*.
- Kate Crawford. 2017. The trouble with bias. In *Conference on Neural Information Processing Systems, invited speaker*.
- Amit Das, Mostafa Rahgouy, Dongji Feng, Zheng Zhang, Tathagata Bhattacharya, Nilanjana Raychawdhary, Mary Sandage, Lauramarie Pope, Gerry Dozier, and Cheryl Seals. 2024. Offlandat: A community based implicit offensive language dataset generated by large language model through prompt engineering. *arXiv preprint arXiv:2403.02472*.
- Sunipa Dev, Emily Sheng, Jieyu Zhao, Aubrie Amstutz, Jiao Sun, Yu Hou, Mattie Sanseverino, Jiin Kim, Akihiro Nishi, Nanyun Peng, et al. 2021. On measures of biases and harms in nlp. *arXiv preprint arXiv:2108.03362*.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- Jwala Dhamala, Tony Sun, Varun Kumar, Satyapriya Krishna, Yada Pruksachatkun, Kai-Wei Chang, and Rahul Gupta. 2021. Bold: Dataset and metrics for measuring biases in open-ended language generation. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 862–872.

- 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*.
- Bosheng Ding, Chengwei Qin, Linlin Liu, Yew Ken Chia, Boyang Li, Shafiq Joty, and Lidong Bing. 2023. [Is GPT-3 a good data annotator?](#) In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11173–11195, Toronto, Canada. Association for Computational Linguistics.
- Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 67–73.
- Soumitra Ghosh, Asif Ekbal, Pushpak Bhattacharyya, Tista Saha, Alka Kumar, and Shikha Srivastava. 2022. Sehc: A benchmark setup to identify online hate speech in english. *IEEE Transactions on Computational Social Systems*, 10(2):760–770.
- Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. Chatgpt outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30):e2305016120.
- Umang Gupta, Jwala Dhamala, Varun Kumar, Apurv Verma, Yada Pruksachatkun, Satyapriya Krishna, Rahul Gupta, Kai-Wei Chang, Greg Ver Steeg, and Aram Galstyan. 2022. Mitigating gender bias in distilled language models via counterfactual role reversal. *arXiv preprint arXiv:2203.12574*.
- Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. 2020. Don’t stop pretraining: Adapt language models to domains and tasks. *arXiv preprint arXiv:2004.10964*.
- Xingwei He, Zhenghao Lin, Yeyun Gong, Alex Jin, Hang Zhang, Chen Lin, Jian Jiao, Siu Ming Yiu, Nan Duan, Weizhu Chen, et al. 2023. Annollm: Making large language models to be better crowdsourced annotators. *arXiv preprint arXiv:2303.16854*.
- Taja Kuzman, Igor Mozetič, and Nikola Ljubešić. 2023. [Chatgpt: Beginning of an end of manual linguistic data annotation? use case of automatic genre identification.](#)
- Alisa Liu, Maarten Sap, Ximing Lu, Swabha Swayamdipta, Chandra Bhagavatula, Noah A Smith, and Yejin Choi. 2021. Dexperts: Decoding-time controlled text generation with experts and anti-experts. *arXiv preprint arXiv:2105.03023*.
- Ioannis Mollas, Zoe Chrysopoulou, Stamatis Karlos, and Grigorios Tsoumakas. 2022. Ethos: a multi-label hate speech detection dataset. *Complex & Intelligent Systems*, 8(6):4663–4678.
- Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.
- Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the fifth international workshop on natural language processing for social media*, pages 1–10.
- Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. 2019. The woman worked as a babysitter: On biases in language generation. *arXiv preprint arXiv:1909.01326*.
- Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. 2020. Towards controllable biases in language generation. *arXiv preprint arXiv:2005.00268*.
- Jiao Sun and Nanyun Peng. 2021. Men are elected, women are married: Events gender bias on wikipedia. *arXiv preprint arXiv:2106.01601*.
- Magdalena Szumilas. 2010. Explaining odds ratios. *Journal of the Canadian academy of child and adolescent psychiatry*, 19(3):227.
- Shuohang Wang, Yang Liu, Yichong Xu, Chenguang Zhu, and Michael Zeng. 2021. [Want to reduce labeling cost? GPT-3 can help.](#) In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4195–4205, Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. *arXiv preprint arXiv:1804.06876*.
- Yiming Zhu, Peixian Zhang, Ehsan-Ul Haq, Pan Hui, and Gareth Tyson. 2023. Can chatgpt reproduce human-generated labels? a study of social computing tasks. *arXiv preprint arXiv:2304.10145*.
- Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2024. Can large language models transform computational social science? *Computational Linguistics*, pages 1–55.

A Appendix

A.1 *HateSpeechCorpus* 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 \$30/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 *HateSpeechCorpus*. 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 4 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 5 presents a selection of sample texts from the *HateSpeechCorpus* dataset accompanied by human-provided annotations.

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, jiggaboo, 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 con, dyke/s, fag, fagbag, fudgepacker, ghey/s, gypo, heeb/s, hilibilly, 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.

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	<i>HateSpeechCorpus</i>	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
I thought kids with Down syndrome were not accepted by families!	ETHOS	Black	Hateful	Not Black	Not Hateful	GPT-4o
As a black person, I think that rappers should stop putting the n-word in the song. Plus, you should be proud of being White. I am proud of being black, so..	ETHOS	Black	Not Hateful	Not Black	Hateful	GPT-3.5
All women love listening to me sperg out, they just hate everything else about me	<i>HateSpeechCorpus</i>	Black	Hateful	Not Black	Not Hateful	GPT-4o
Sweet and cute yet will cripple a beta with a snap of a finger.	<i>HateSpeechCorpus</i>	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
Syria has a religion problem..	ETHOS	Muslim	Hateful	Not Muslim	Not Hateful	GPT-4o
I see the depression caused her to gain more weight.	ETHOS	Physical Disability	Hateful	No Disability	Not Hateful	GPT-4o
Is this guy actually autistic or Is he faking it	ETHOS	Mental Disability	Hateful	No Disability	Not Hateful	GPT-4o
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
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
Eliminate their infrastructure and cripple them.	<i>HateSpeechCorpus</i>	Muslim	Hateful	Not Muslim	Not Hateful	GPT-4o
Like there was that one time where he was the gimp in the Armageddon and that other time, Magic Mike or whatever	<i>HateSpeechCorpus</i>	Physical Disability	Hateful	No Disability	Not Hateful	GPT-4o

Continued on next page

Text	Dataset	Annotator	Annotation	Annotator	Annotation	LLM
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	<i>HateSpeechCorpus</i>	Mental Disability	Hateful	No Disability	Not Hateful	GPT-4o
Houston is an important petro chemical hub. Bomb Houston and you could cripple multiple industries.	<i>HateSpeechCorpus</i>	Muslim	Not Hateful	No Muslim	Hateful	GPT-3.5
It's scary how "Technical Debt Cripples Companies And Threatens To Stifle Innovation" - learn more in this segment on with vFunction CEO	<i>HateSpeechCorpus</i>	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.	<i>HateSpeechCorpus</i>	Mental Disability	Not Hateful	No Disability	Hateful	GPT-3.5
Houston is an important petro chemical hub. Bomb Houston and you could cripple multiple industries.	<i>HateSpeechCorpus</i>	Muslim	Not Hateful	No Muslim	Hateful	GPT-3.5
It's scary how "Technical Debt Cripples Companies And Threatens To Stifle Innovation" - learn more in this segment on with vFunction CEO	<i>HateSpeechCorpus</i>	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.	<i>HateSpeechCorpus</i>	Mental Disability	Not Hateful	No Disability	Hateful	GPT-3.5

Table 4: Sample Texts with annotation biases while annotating by GPT-3.5 and GPT-4o.

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 5: Sample texts from *HateSpeechCorpus* with human annotation.