Proceedings of the Nineteenth International AAAI Conference on Web and Social Media (ICWSM 2025)

Large Language Model Annotation Bias in Hate Speech Detection

Ebuka Okpala, Long Cheng

School of Computing
Clemson University
{eokpala, lcheng2}@clemson.edu

Abstract

Large language models (LLMs) are fast becoming ubiquitous and have shown impressive performance in various natural language processing (NLP) tasks. Annotating data for downstream applications is a resource-intensive task in NLP. Recently, the use of LLMs as a cost-effective data annotator for annotating data used to train other models or as an assistive tool has been explored. Yet, little is known regarding the societal implications of using LLMs for data annotation. In this work, focusing on hate speech detection, we investigate how using LLMs such as GPT-4 and Llama-3 for hate speech detection can lead to different performances for different text dialects and racial bias in online hate detection classifiers. We used LLMs to predict hate speech in seven hate speech datasets and trained classifiers on the LLM annotations of each dataset. Using tweets written in African-American English (AAE) and Standard American English (SAE), we show that classifiers trained on LLM annotations assign tweets written in AAE to negative classes (e.g., hate, offensive, abuse, racism, etc.) at a higher rate than tweets written in SAE and that the classifiers have a higher false positive rate towards AAE tweets. We explore the effect of incorporating dialect priming in the prompting techniques used in prediction, showing that introducing dialect increases the rate at which AAE tweets are assigned to negative classes.

1 Introduction

Large language models (LLMs) require large pretraining datasets obtained from crawling the internet (Radford et al. 2019; Raffel et al. 2020) to learn world knowledge and prevent overfitting. However, the problem with such datasets is that they contain texts that exhibit biases or stereotypes observed in our society, which has negative implications when models trained on such data are used in real-world applications such as in hate speech or toxicity detection (Gehman et al. 2020).

One such implication is racial bias, where the models discriminate against texts written in African American English (AAE) at a higher rate than texts written in Standard American English (SAE). In an online social platform such as X (Twitter), such a model would over moderate tweets written in AAE failing to protect the group they were designed to protect and could reduce the voices of marginalized groups during global social movements (Tyler and Smith 1995). AAE is a dialect of American English with defined syntactic-semantic, phonological, and lexical features (Blodgett and O'Connor 2017).

As LLMs become ubiquitous, for example, GPT-like models such as ChatGPT (Schulman et al. 2022), their popularity indicates their potential to be utilized in different applications. Researchers have explored the use of GPT-3 as a low-cost data labeler (Wang et al. 2021), leveraged ChatGPT for the annotation of implicit hate speech (Huang, Kwak, and An 2023), and studied the limitations of using ChatGPT for data annotation (Thapa, Naseem, and Nasim 2023). The use of GPT-3 annotated data to train downstream models has been explored by researchers (Wang et al. 2021); despite the success of LLMs, they have a high loss in quality compared to state-of-the-art (SOTA) methods in difficult and pragmatic tasks (Kocoń et al. 2023) such as aggression and especially for emotion classification task.

In this work, we systematically analyze how LLMs can perform differently for different dialects and propagate racial bias in downstream models trained on the LLM annotated data. We focus on evaluating hate speech detection classifiers trained on GPT-4 annotated datasets and generalize this analysis to Llama-3. We show that utilizing LLMs for data annotation can introduce racial bias in downstream models. Our research aims to help maintain civility in conversation on social platforms while highlighting the benefits and, importantly, the risks of using LLMs in annotating data for hate speech detection. We summarize our main **contributions below**[1]:

- We use GPT-4 to predict hate speech in seven hate speech detection datasets collected from X (Twitter) using three prompting techniques (general, few-shot learning, and chain-of-thought reasoning). We compare the performance of GPT-4 predictions to the predictions of classifiers trained on human-annotated hate speech datasets. Evaluations are performed when datasets are conditioned on dialect (AAE and SAE) and when they are not.
- We fine-tuned three pre-trained language models often used in the hate speech literature on the datasets re-annotated under each prompting technique and measured racial bias in each resulting classifier. Specifically, we

Copyright © 2025, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

[1] code: https://github.com/burunkus/llm_annotation_bias

focus on the racial disparity between text written in AAE and SAE in classifiers trained on GPT-4 annotated datasets.

- We evaluate racial bias using a corpus of demographically aligned tweets to show how each classifier performs on AAE and SAE tweets and AUC-based metrics to calculate the false positive rates of each classifier on the test sets conditioned on dialect.
- We show that the problem of racial bias generalizes to other LLMs, specifically Llama-3.

Extensive evaluation of 63 (seven datasets, three models, and three prompting techniques) classifiers shows evidence of racial bias across all the classifiers and prompting techniques, with AAE tweets assigned to negative classes at a higher rate than SAE tweets and the classifiers having more false positives for AAE tweets than SAE tweets. Compared to classifiers trained on human-annotated data, classifiers trained on LLM-annotated data can increase the rate of classifying AAE tweets to negative classes. We expect that if LLMs are used for data annotation and subsequently to train downstream classifiers deployed in the field, the classifiers will discriminate against those who write in AAE, most who are African-American known to experience racial discrimination in a wide range of applications such as in housing (Massey and Lundy 2001) and criminal justice (Rickford and King 2016) which feeds into the ideology and stereotypes about African-Americans (Bergsieker et al. 2012; Ghavami and Peplau 2013).

2 Related Work

Racial Bias and Toxicity in Language Models Past work has studied racial bias towards AAE tweets in machine learning (Davidson, Bhattacharya, and Weber 2019) and introduced various methods for the mitigation of racial bias in deep learning models, from adversarial debiasing in traditional deep learning (Xia, Field, and Tsvetkov 2020) and transformer-based models (Okpala et al. 2022), to regularization based techniques (Mozafari, Farahbakhsh, and Crespi 2020). More recently, (Hofmann et al. 2024) demonstrated dialect prejudice in LLMs using matched guise probing. Results indicate that speakers of AAE are more likely to be assigned less attractive jobs, be convicted of crimes, and be sentenced to death by LLMs due to the features of AAE than speakers of SAE. Previous studies have also explored the generation of toxic texts by LLMs, (Gehman et al. 2020) introduced the RealToxicityPrompt dataset of sentence prefixes paired with their toxicity score from Perspective API[2]. They showed that pre-trained autoregressive language models can be prompted to generate toxic text even with non-toxic prompts. Using the same dataset, (Schick, Udupa, and Schütze 2021) focusing on the generation of biased text by GPT-2 and T5 (Raffel et al. 2020) demonstrated that language models are aware of their biases and the toxicity of the text they generated.

Hate Speech Data Annotation Using Large Language Models Researchers have evaluated the performance of LLMs as annotators for various NLP tasks. GPT-3 was used with different annotation strategies for annotating multiple NLP tasks, from sentiment analysis to named entity recognition, by fine-tuning BERT_base model on the GPT-annotated data (Ding et al. 2023). ChatGPT has been explored as an assistive tool during annotation (Mei et al. 2023), as the sole annotator (Gilardi, Alizadeh, and Kubli 2023; Huang, Kwak, and An 2023), and in the annotation of sentiments towards volitional entities in long texts (Rønningstad, Velldal, and Øvrelid 2024). The study by (Wang et al. 2021) indicates that directly using GPT-3 for downstream tasks may not produce the best performance when compared to using downstream models such as PEGASUS_large or RoBERTa_large fine-tuned on GPT-3 annotated datasets for downstream tasks.

Hate Speech Detection The problem of hate speech in online social platforms as a critical threat (Thomas et al. 2021) has been tackled by researchers using traditional machine learning approaches (Schmidt and Wiegand 2017; Fortuna and Nunes 2018), deep neural network methods (Badjatiya et al. 2017; Maity et al. 2024), transformer-based approaches such as the use of BERT in (Liu, Li, and Zou 2019; Guo et al. 2023), COVID-Twitter-BERT in (Liao et al. 2023), and the retraining of BERT_base on COVID-19 related hateful tweets and on posts from banned Reddit subcommunities to produce COVID-HateBERT (Li et al. 2021) and HateBERT (Caselli et al. 2021) respectively. Most recently, the capabilities of LLMs have been explored in hate speech detection (Guo et al. 2023; Vishwamitra et al. 2024; He et al. 2023; Kocoń et al. 2023; Li et al. 2024).

In contrast to our work, all these ideas focus on evaluating models trained on human-annotated hate speech datasets for racial bias or evaluating socially undesirable attributes or biases in text generated by generative models. The works on data annotation explored the effectiveness of GPT in classifying implicit hate, the possibility of GPT-like LLMs replacing human annotators, and the implications of using GPT-like LLMs for data annotation. While they discussed the advantages and disadvantages of using GPT, the empirical evidence of its negative implications is lacking in terms of racial bias. Our work differs from these works by qualitatively showing evidence of racial bias in using LLMs such as GPT-4 and Llama-3 to annotate data used in downstream models, specifically in hate speech detection.

3 Methodology

This section details our methodology for assessing bias in downstream models trained on LLM-annotated hate speech detection datasets.

3.1 Data

We utilize two types of datasets in our study, as shown in Fig 1. The race dataset described below is used to extract datasets used in training an AAE language model and a dialect classifier. The hate speech datasets described below are used for training hate speech classifiers.

Race Dataset Following (Okpala et al. 2022; Mozafari, Farahbakhsh, and Crespi 2020; Davidson, Bhattacharya, and

[2]https://github.com/conversationai/perspectiveapi

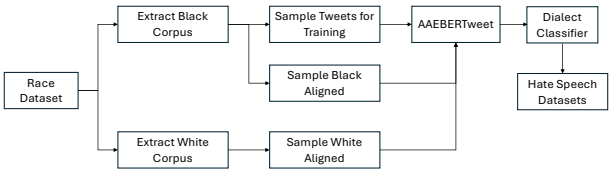


Figure 1: An overview of the white-aligned and black-aligned datasets creation from the race dataset.

Weber 2019), we use a race dataset introduced by (Blodgett, Green, and O'Connor 2016) to measure racial bias in classifiers trained on LLM annotated datasets and to train a dialect classifier. Blodgett et al. (Blodgett, Green, and O'Connor 2016) collected 59.2M tweets mapped according to the location the tweet author lived in using the geo-location published by the tweet author. They matched each tweet to the US Census block group they were sent in, and each user was mapped as non-Hispanic whites, non-Hispanic blacks, Hispanics, and Asians using the block group race and ethnicity information. They trained a mixed-method probabilistic model that learns demographically-aligned language models for each demographic. Each model calculates the posterior probability of using a language in a tweet. We follow the authors in (Okpala et al. 2022; Mozafari, Farahbakhsh, and Crespi 2020) and extract a black-and-white corpus containing tweets with a posterior probability > 0.8. We obtained a black-and-white corpus of 1.28M and 19M tweets, respectively. From the black corpus, we randomly sample two sets of tweets. The first set contains 1.22M tweets used to train an AAE language model (Okpala et al. 2022) as described in Section 3.2; from the remaining set of tweets, we randomly sample the second set containing 1K tweets which we call *black-aligned* tweets used to fine-tune a dialect classifier described in Section 3.2 and for racial bias assessment. We randomly sample 1K tweets from the white corpus, which we call *white-aligned* tweets also used to fine-tune a dialect classifier and for racial bias assessment. Fig. 1 shows the data extraction process of the white-aligned and *black-aligned* tweets.

Hate Speech Datasets To understand the racial bias in hate speech detection models trained on LLM annotated data, we focus our analyses on seven corpora of tweets (Waseem 2016; Davidson et al. 2017; Golbeck et al. 2017; Founta et al. 2018; Zampieri et al. 2019; Caselli et al. 2020; Basile et al. 2019) widely used in hate speech detection (Park, Shin, and Fung 2018; Lee, Yoon, and Jung 2018; Van Aken et al. 2018; Kapoor et al. 2019; Caselli et al. 2021; Li et al. 2021) and racial bias assessment (Davidson, Bhattacharya, and Weber 2019; Sap et al. 2019; Mozafari, Farahbakhsh, and Crespi 2020; Okpala et al. 2022). We utilized random samples of these datasets in our experiments due to the cost of annotating large samples with OpenAI's GPT-4 using different prompting strategies for annotation discussed in Section 3.3. If the number of tweets in each dataset class exceeds 700, we randomly sampled 500 tweets from the class. Otherwise, we retained the original number of tweets in the class. The statistics of these datasets are shown in Table 1[3], a detailed statistics is shown in Table 8 of the Appendix. Following (Okpala et al. 2022), we don't utilize the spam label in the Founta (Founta et al. 2018) dataset in our analysis. Details of each dataset can be found in the referenced dataset paper; a summary of the datasets can be found in (Okpala et al. 2022).

Dataset	Count	Count after sampling
Waseem (Waseem 2016)	5,988	1,116
Davidson (Davidson et al. 2017)	24,773	1,500
Founta (Founta et al. 2018)	45,549	1,500
Golbeck (Golbeck et al. 2017)	20,305	1,000
OffensEval (Zampieri et al. 2019)	14,100	1,860
AbusEval (Caselli et al. 2020)	14,100	2,360
HatEval (Basile et al. 2019)	11,991	2,000

Table 1: Statistics of the datasets.

3.2 AAE Language Model and Dialect Classifier

The authors in (Okpala et al. 2022) introduced AAEBERT, an African-American English language model. AAEBERT was obtained by retraining BERT (Devlin et al. 2018) on a subset of tweets written by non-Hispanic Blacks from the race dataset (Blodgett, Green, and O'Connor 2016). While BERT (Devlin et al. 2018) have shown improved performance on several NLP tasks, BERTweet (Nguyen, Vu, and Nguyen 2020), pre-trained on English Twitter data have shown better performance on several Tweet NLP tasks. Due to the performance of BERTweet (Nguyen, Vu, and Nguyen 2020) on English tweets and the fact that our language of interest (AAE) is a variation of English, BERTweet will more likely capture the nuances of AAE than BERT (Devlin et al. 2018). In this work, we reproduce AAEBERT (Okpala et al. 2022) by retraining BERTweet (Nguyen, Vu, and Nguyen 2020) on the 1.22M black corpus extracted from the race dataset as discussed in Section 3.1 and call this retrained model ***AAEBERTweet***. We follow the implementation details of AAEBERT as described in (Okpala et al. 2022) using masked language modeling as the training objective, a maximum sequence length of 100, batch size of 64, and training for 100 epochs on one V100 GPU.

We train a dialect classifier to infer the dialect of each tweet in the hate speech datasets described in Section 3.1 because a race label is required to assess racial bias using the AUC metrics described in Section 3.5. The authors in (Okpala et al. 2022) directly used the AAEBERT language model with sigmoid activation function to classify a tweet as AAE with a threshold > 0.5 and as SAE otherwise. Contrary to this, we fine-tune AAEBERTweet on the 1K black-aligned and 1K white-aligned tweets to obtain a

[3]After tweet rehydration, the total count of tweets in some datasets does not sum to the count originally published because some tweets have been removed by X (Twitter).

Target	F1	Precision	Recall
AAE	0.846	0.852	0.839
SAE	0.849	0.843	0.856

Table 2: Performance of the dialect classification model for the AAE and non-AAE (SAE) classes. Evaluation metrics are macro averages.

dialect model used to infer dialects as shown in Fig 1. The fine-tuned model was used to classify the tweets in each hate speech dataset discussed in Section 3.1 as AAE and SAE. The model trained with a learning rate of $1e-5$, batch size of 32, and for 20 epochs achieved 0.848, 0.847, and 0.847 precision, recall, and F1 scores, respectively. To show that our dialect classifier obtained by fine-tuning AAEBERTweet performs better than AAEBERT (Okpala et al. 2022), we retrain BERT to obtain AAEBERT as described in (Okpala et al. 2022), then fine-tuned AAEBERT on the 1K black-aligned and 1K white-aligned tweets. The resulting dialect model achieved 0.800, 0.800, and 0.800 precision, recall, and F1 scores, a less-performing dialect model compared to our dialect model obtained from fine-tuning AAEBERTweet. The per-class performance of our dialect model is shown in Table 2. We preprocess each tweet in the black-and-white-aligned tweets by replacing hyperlinks with HTTPURL, removing the # sign in hashtags, replacing handles with @USER, replacing extra white space with single space, replacing numbers with NUMBER, removing punctuations, and ensuring each tweet contained more than three words. The evaluation of the dialect classifier is discussed in Section A.1 of the Appendix.

3.3 Prompt Annotation

We employ various prompting strategies for data annotation to determine bias in classifiers trained on LLM-annotated datasets. We used the GPT-4-0613 model from the official OpenAI API endpoints to run the various prompts for annotating each dataset. For generalization, we used Meta's Llama-3-8B-Instruct model (Meta 2024) on HuggingFace to annotate each dataset. We used a variation of prompting strategy utilized in hate speech detection and annotation (Guo et al. 2023; Huang, Kwak, and An 2023; Li et al. 2024). Each prompting strategy is described below with examples in Table 9 of the Appendix.

General Prompt Annotation The general (Gen) prompt technique allows the adaptation of LLMs for the specific task of hate speech annotation. Given a tweet x, the input to the LLM in this annotation strategy is formatted as: *Given the tweet in triple quotes: """x""". Do you think the tweet is [classes]? Only answer with one of the following: [classes]. Do not provide an explanation for your answer.* Where [classes] represent the original classes or categories in each human-annotated dataset, for example, in the Davidson dataset, [classes] = hate or offensive or normal. The LLM will then output y, for example, either "hate", "offensive" or "normal" representing the annotation for x for the Davidson dataset. We use the general prompt technique as it is likely the most common way general users can perform annotation in the wild. We complement this technique with the techniques described below to offer a broader perspective on racial bias.

Few-shot Prompt Annotation Few-shot (FS) learning has improved LLMs' performance in many NLP tasks (Brown et al. 2020). In this annotation setting, examples, also known as few-shot demonstrations with answers or solutions, are included in the prompt given to an LLM, essentially demonstrating the task to the LLM; the LLM learns in context via prompting. We explore whether racial bias persists in this annotation setting as it is likely that data annotation can be performed in the real world using an LLM with a few labeled examples to improve annotation and to avoid over-exposure to hateful content. We randomly sampled two exemplars from each class in each dataset, which were used as part of the prompt to assess racial bias in downstream models trained on datasets annotated with few-shot demonstrations. One of the exemplars is detailed in Table 9 of the Appendix.

Chain-of-Thought Prompt Annotation Finally, we explored Chain-of-Thought (CoT) prompting annotation, a series of intermediate natural language reasoning steps that lead to the final answer (Wei et al. 2022). The Chain-of-Thought prompt has been shown to enhance the ability of LLMs to solve complex reasoning tasks in NLP (Wei et al. 2022), and it consists of triples: [x, chain of thought, y]. We modify the few-shot exemplars in the few-shot prompt annotation setting for CoT annotation as shown in Table 9 of the Appendix. The intermediate natural language reasoning steps are designed by elaborating the definitions of hate speech as defined by the authors of each dataset. Each example used in the few-shot annotation setting is augmented with an answer with comprehensive reasoning to explain why the example belongs to a particular class or category. As in few-shot prompt annotation, CoT reasoning is likely to be used in the real world for data annotation, where a few examples are provided together with a reasoned explanation of why a text is hateful or not hateful. We simulate that scenario in this setting.

We tested different variations of these prompts and settled for the stated prompts because they worked and exhibited good performance across all the seven datasets analyzed and across different numbers of class labels. We note that for our generalization experiment using Llama-3, we only analyzed datasets annotated using general prompt annotation due to the difficulty in finding a prompt that works across LLMs (specifically GPT-4), datasets, and classes for FS and CoT. To maintain consistent results, we skip FS and CoT annotation with Llama-3.. As discussed in Section 3.1, we sampled from the hate speech datasets because of the cost associated with annotating using FS and CoT prompting techniques via Open AI as charges are token-based[4].

[4]https://help.openai.com/en/articles/7102672-how-can-i-access-gpt-4-gpt-4-turbo-and-gpt-4o

3.4 Hate Speech Classifiers

For each LLM annotated dataset, we train a classifier on the training dataset to predict the class of each tweet in the test dataset. For the datasets not initially split into train and test sets by the original authors, we randomly split those datasets (Waseem, Davidson, Founta, and Golbeck) into train and test sets using the 80:20 splits. We use the same set of pre-trained models used in (Okpala et al. 2022), BERT (Devlin et al. 2018) (bert-base-uncased on HuggingFace), BERTweet (Nguyen, Vu, and Nguyen 2020) (vinai/betweet-base on HuggingFace), and HateBERT (Caselli et al. 2021). We fine-tuned the pre-trained models on the LLM-annotated datasets to obtain twenty-one hate speech classifiers (seven datasets on three pre-trained models) for each annotation strategy used. We fine-tuned the pre-trained models on human-annotated datasets for comparison purposes and obtained twenty-one hate speech classifiers. Each classifier was trained using a learning rate of $1e-5$, a batch size of 32, a maximum sequence length of 100, 5 epochs, an Adam optimizer, and cross-entropy loss. We used the same pre-processing steps used in Section 3.2 to pre-process each tweet, except filtering tweets that do not have at least four words.

3.5 Bias Metrics

We use the evaluation metrics described below to evaluate bias in classifiers trained on LLM-annotated datasets annotated using different prompting strategies.

Hypothesis-based Metric The hypothesis-based evaluation metric (Davidson, Bhattacharya, and Weber 2019; Mozafari, Farahbakhsh, and Crespi 2020; Okpala et al. 2022) assesses whether the difference in the probability of a tweet being predicted as a particular class is due to the tweet author's race[5]. We use the dialect of the tweet as a proxy for race (Davidson, Bhattacharya, and Weber 2019; Mozafari, Farahbakhsh, and Crespi 2020; Okpala et al. 2022). The evaluation is based on estimating the proportion of tweets in each dataset that each classifier classifies as belonging to each class using the sampled black-aligned and white-aligned tweets discussed in Section 3.1. We define a null hypothesis (H_N) that there is no racial bias if the probability of a tweet belonging to a negative class is independent of the author's race. We test $H_N : P(c_i = 1|black) = P(c_i = 1|white)$ for each negative class c_i, where $c_i = 1$ represents membership in the class and $c_i = 0$ represents otherwise. We reject the null hypothesis H_N in favor of the alternative H_A that black-aligned tweets are classified to negative classes c_i at a higher rate than white-aligned tweets if $P(c_i = 1|black) > P(c_i = 1|white)$ and the difference is statistically significant. If $P(c_i = 1|black) < P(c_i = 1|white)$, then white-aligned tweets are assigned to negative classes at a higher rate. We create a vector per class for each racial group (black and white) in which each element is the probability p_i of a tweet belonging to a negative class i as predicted by a classifier. We obtain vectors of dimension $n = 1000$ (the number of tweets in the black-aligned and white-aligned datasets). For each group, we calculate the proportion of tweets assigned to negative class i as $\widehat{p_{iblack}} = \frac{1}{n}\sum_{j=1}^{n} p_{ij}$ for black-aligned and $\widehat{p_{iwhite}} = \frac{1}{n}\sum_{j=1}^{n} p_{ij}$ for white-aligned. We test $\widehat{p_{iblack}} = \widehat{p_{iwhite}}$ for significance using t-test. If the magnitude of the difference $\frac{\widehat{p_{iblack}}}{\widehat{p_{iwhite}}} > 1$, then black-aligned tweets are assigned to a negative class at a higher rate than white-aligned.

AUC-based Metric Machine learning classifiers can exhibit unintended bias as the systemic differences in performance for different demographic groups (Borkan et al. 2019). The AUC-based metrics introduced by the Google Conversational AI Team (Borkan et al. 2019) have been used to measure identity-based (such as "gay", "muslim", etc.) unintended bias in machine learning classifiers for hate speech detection (Vaidya, Mai, and Ning 2020; Mathew et al. 2021). We used AUC-based metrics described below to assess racial bias towards tweets written in AAE and SAE by classifiers trained on LLM-annotated datasets. We focus on the ability of the classifiers to reduce false positive rates on non-hateful tweets inferred to be written in AAE known empirically to introduce model bias. The AUC metrics include Subgroup AUC, Background Positive Subgroup Negative (BPSN), and Generalized Mean of Bias AUCs. For these metrics, we convert datasets with multi-class LLM-annotated labels into binary labels, re-trained our classifiers on the binary labels (hate and non-hate) and evaluate the reduction of unintended bias towards a group by the classifiers. Evaluation is restricted to the test set of each dataset and not on the black-aligned and white-aligned datasets to understand how the classifiers perform in hate speech detection and bias reduction.

Subgroup AUC: We restrict the test set to hateful and non-hateful tweets written in AAE and SAE. The ROC-AUC score is calculated for each group (AAE and SAE), resulting in the Subgroup AUC for a group. This metric measures the model's ability to separate hateful and not hateful tweets in the context of a specific group. A higher score indicates that the classifier is doing an excellent job of separating hateful and non-hateful posts particular to the racial group.

BPSN (Background Positive, Subgroup Negative AUC): We restrict the test set to non-hateful tweets written in AAE and hateful tweets not in AAE. The BPSN AUC is obtained by calculating the ROC-AUC score of this set. The false positive rate of each classifier in the context of each specific group is measured by this metric. A classifier is less likely to confuse non-hateful tweets written by a group with hateful tweets not written by the group if the BPSN score is high, meaning the model can reduce bias towards a specific group. We consider BPSN a stronger metric than Subgroup AUC because it aligns with the focus of this paper, which is the false positive rate towards certain groups.

Generalized Mean of Bias AUCs: As part of their Kaggle competition[6], the Google Conversational AI Team intro-

[5]Blodgett et al.(Blodgett, Green, and O'Connor 2016) show that although not all AAs speak AAE, the use of AAE dialect suggests a social proximity to or affinity for African American communities

[6]https://www.kaggle.com/c/jigsaw-unintended-bias-in-toxicity-classification/overview/evaluation

duced this metric which combines the per-group Bias AUCs into an overall measure as $M_p(m_s) = (\frac{1}{n}\sum_{s=1}^{N} m_s^p)^{\frac{1}{p}}$ where, M_p is the p^{th} power-mean function, m_s is the bias metric calculated for subgroup s, N is the number of groups which is 2, and p is set to -5 as done in the competition. We report the following metrics for our datasets:

- **GMB-Subgroup-AUC**: GMB AUC with Subgroup AUC as the bias metric
- **GMB-BPSN-AUC**: GMB AUC with BPSN AUC as the bias metric

4 Results

In this section, we discuss the results of our study examining racial bias in using LLMs for data annotation in hate speech detection. We discuss results in two settings: per dataset and combined datasets. In the combined datasets setting, we combined the seven datasets annotated using LLM general prompt annotation into one dataset having a binary class, which was used to train our classifiers. We combine datasets in general prompt annotation because the prompt scales across LLMs, enabling a more informed comparison between LLMs.

Performance In the combined datasets setting, as shown in Table 3, the combined GPT-4-annotated datasets outperform both the combined Llama-3 and human-annotated datasets across classifiers and metrics. The BERTweet classifier outperforms other classifiers across metrics. In the per-dataset setting, for the Llama-3 annotated datasets as shown in Tables 11 and 10 of the Appendix, BERTweet performs the best in six of the datasets in binary classification and is competitive in multi-class classification.

For the GPT-4 annotated datasets, the overall binary and multi-class classification performance for various prompting strategies is summarized in Tables 21 and 12 of the Appendix. From the multi-class classification results in Table 12, we observe that for general prompt annotation, BERTweet is competitive, for FS and CoT, HateBERT and BERTweet outperform almost all models, respectively. Overall, the use of FS learning prompt annotation increases performance. For the binary classification in Table 21, BERTweet outperforms other models in five and six datasets across the prompt annotation strategies, respectively. Similar to multi-class classification, FS prompt annotation increases performance consistently across models and datasets except in the BERTweet model fine-tuned on the AbusEval dataset and in the BERT and BERTweet models fine-tuned on the HatEval datasets. When compared to the model performance on human-annotated datasets as shown in Tables 13 and 14 in the Appendix for multi-class and binary label classification, models fine-tuned on GPT-4 annotated datasets outperforms models fine-tuned on human-annotated datasets in binary classification in 3 datasets for gen, 4 datasets for FS, and 4 datasets for CoT annotation. In multi-class classification, models fine-tuned on human-annotated datasets perform better (in 3 datasets, Davidson, Founta, and AbusEval) than models fine-tuned on datasets annotated using general prompt annotation. The overall performance of GPT-4 general prompt annotation and human-annotation are shown in Table 15 of the Appendix, and the performance when each dataset is conditioned on each dialect is shown in Table 16 of the Appendix. The FS and CoT prompt annotation overall performance results are shown in Tables 17 and 18 of the Appendix, respectively, and the performance results when the datasets are conditioned on dialect are shown in Tables 19 and 20 of the Appendix. When the performance of FS and CoT annotation is compared to the general prompt annotation results in Tables 15 and 16 of the Appendix, we observe that the level of agreement between GPT-4 annotation and human annotation increases across almost all datasets with or without conditioning on dialect and across nearly all metrics.

Finally, we use Cohen's Kappa to compare LLM-generated and human annotations. Table 4 shows the κ scores of the LLM annotation of each dataset compared to human annotations. For GPT-4, we observe that the level of agreement increases in FS and CoT annotations. The highest level of agreement is observed in the CoT annotation of the Waseem dataset (72.4%), and the lowest is observed in the general prompt annotation of the Founta dataset (18.7%). Except for the Waseem and Hateval datasets, the quality of GPT-4 generated labels is poor compared to human labels. A similar observation is made for Llama-3-generated labels when compared to human labels. Table 35 of the Appendix shows the poor κ scores of LLM annotations with dialect priming compared to human annotation. We discuss dialect priming in Section A.2 of the Appendix.

Bias - Hypothesis based Tables 5 and 6 show the results of the models trained in the combined datasets setting for GPT-4 and Llama-3 general prompt annotation. From Table 5, there are statistically significant differences ($p \ll 0.05$) in how all classifiers assign black-aligned tweets into negative classes at a higher rate than white-aligned tweets for both human and GPT-4 annotations. The BERTweet classifier shows the highest disparity by frequently assigning black-aligned tweets as hate at 2.2 times for human annotation and 2.7 and 2.2 times for GPT-4 and Llama-3 annotations, respectively. Classifiers trained on the combined GPT-4 annotated datasets are more biased towards AAE than classifiers trained on the combined Llama-3 annotated datasets.

In the per dataset setting, table 23 of the Appendix shows the results of the BERT model fine-tuned on datasets annotated by GPT-4 using general prompt annotation. From Table 23, we observe racial disparities in the performance of most of the classifiers. There are statistically significant differences ($p \ll 0.05$) in most classifiers except in two instances. In all the statistically significant instances, we observe that black-aligned tweets are assigned to negative classes at a higher rate than white-aligned tweets except for one instance, in the implicit class of the AbusEval classifier where black-aligned tweets are assigned to a negative class at a lower rate than white-aligned tweets. For the Davidson and Waseem classifiers, we observe no significant difference in the rates at which tweets are classified as hate and racism, respectively, with the rates remaining low. Black-aligned tweets are classified frequently as offensive at 1.4

Dataset	Model	GPT-4			Llama-3			Human		
		P	R	F1	P	R	F1	P	R	F1
All	BERT	0.799	0.805	0.801	0.782	0.779	0.779	0.733	0.734	0.713
	BERTweet	**0.835**	**0.839**	**0.837**	**0.814**	**0.811**	**0.812**	**0.746**	**0.747**	**0.727**
	HateBERT	0.809	0.812	0.810	0.791	0.787	0.787	0.734	0.736	0.719

Table 3: Classifier performance after fine-tuning on the single dataset obtained by combining the seven datasets annotated using general prompt and human annotation. The combined GPT-4 annotated datasets and BERTweet classifier outperforms the combined Llama-3 and human annotated datasets and classifiers across all metrics (Precision (P), Recall (R), and F1). Evaluation metrics are macro averages.

Dataset	GPT-4			Llama-3
	Gen	FS	CoT	Gen
Davidson	0.416	0.625	0.639	0.294
Founta	0.187	0.263	0.399	0.274
AbusEval	0.373	0.407	0.448	0.352
Waseem	0.617	0.727	0.724	0.512
Golbeck	0.292	0.342	0.324	0.236
OffensEval	0.473	0.524	0.454	0.404
HatEval	0.501	0.521	0.565	0.448

Table 4: The Cohen's Kappa scores between GPT-4 annotated datasets using various prompting strategies and human annotation and Llama-3 annotated datasets using general prompt annotation and human annotation.

times, hate at 1.7 times, and abuse at 1.6 times compared to white-aligned tweets in the OffensEval, HatEval, Davidson, and Founta classifiers. Similar observation on racial disparities is made in the BERT model fine-tuned on GPT-4 annotated datasets using FS and CoT prompt annotation as shown in Table 29 of the Appendix. Comparing the human annotation results in Table 23 with the general, FS, and CoT annotation results in the Tables 23 and 29 (See Appendix), we see that there are instances ($p \ll 0.05$), OffensEval, HatEval, Founta, Golbeck, offensive class of Davidson using general annotation, and in the sexism class of Waseem using CoT annotation, where GPT-4 annotation increases racial disparity. There are statistically significant ($p \ll 0.05$) instances, explicit class of AbusEval, sexism class of Waseem using general and FS annotation, and offensive class of Davidson using FS and CoT annotation, where GPT-4 annotation reduces the rate of assigning black-aligned tweets to negative classes. A change in direction occurs in the hate class of Davidson using FS and CoT annotation and the racism and sexism class of Waseem ($p \ll 0.05$).

The results of the BERTweet model fine-tuned on datasets annotated using general prompt annotation are shown in Table 25 of the Appendix, and the results of using FS and CoT prompt annotation are shown in Table 30 of the Appendix. From Table 25, we see results similar to the BERT model. Comparing the human annotation results with the general, FS, and CoT annotation results as shown in Tables 25 and 30 (See Appendix), there are increases in the rate at which black-aligned tweets are classified as negative classes more than white-aligned tweets in some statistically significant ($p < 0.05$) instances, OffensEval, HatEval, Golbeck, abuse class of Founta, hate class of Founta using FS and CoT annotation, implicit class of AbusEval, and racism and sexism class of Waseem. There are instances ($p < 0.05$), offensive class of Davidson, hate class of Davison using FS and CoT annotation, hate class of Founta using general annotation, explicit class of AbusEval using general annotation, sexism class of Waseem, and racism class of Waseem using general annotation, where there are reductions in the rate of assigning black-aligned tweets to negative classes using prompt annotations. A significant ($p < 0.05$) change in direction is observed in the hate class of the Davidson classifier fine-tuned on the Davidson dataset annotated using general prompting annotation.

The results of the HateBERT model obtained by fine-tuning datasets annotated using general prompt annotation as shown in Table 27 and FS and CoT prompt annotation as shown in Table 31 of the Appendix are consistent with the BERT and BERTweet results. When compared to the classifiers fine-tuned on human annotated datasets focusing on significant ($p \ll 0.05$) instances, the rate of classifying black-aligned tweets to negative classes increases in AbusEval, OffensEval, HatEval, Founta, and Golbeck classifiers and the racism class of Waseem and the sexism class of Waseem using CoT annotation. There are a few decreases in the offensive class of Davidson, the hate class of Davidson using CoT, the sexism class of Wasee using FS, and a change in direction in the hate class of Davidson using general prompt annotation.

The results of fine-tuning various classifiers on the Llama-3 annotated datasets using general prompt annotation are in line with the results of GPT-4 general prompt annotation as shown in Tables 24, 26, and 28 of the Appendix. Just like GPT-4, Llama-3 exhibits significant differences in racial disparity between AAE and SAE tweets by assigning AAE tweets to negative classes at a higher rate than SAE tweets, and there are instances where there is a change in direction in racial disparity. In summary, these results demonstrate that using an LLMs, for data annotation can introduce racial bias in the annotated dataset. If the biased dataset is used to fine-tune a model for downstream tasks, as we have seen in hate speech detection, the downstream models will propagate the racial bias introduced by the LLM. We have shown that tweets written in AAE are disproportionately assigned

Class	Model	Human annotated					GPT-4 annotated (general prompt)				
		$\widehat{p_{iblack}}$	$\widehat{p_{iwhite}}$	t	p	$\frac{\widehat{p_{iblack}}}{\widehat{p_{iwhite}}}$	$\widehat{p_{iblack}}$	$\widehat{p_{iwhite}}$	t	p	$\frac{\widehat{p_{iblack}}}{\widehat{p_{iwhite}}}$
Hate	BERT	0.431	0.229	13.204	***	1.878	0.287	0.128	11.243	***	2.245
	BERTweet	0.379	0.174	14.415	***	2.185	0.238	0.089	11.446	***	2.691
	HateBERT	0.406	0.229	12.554	***	1.768	0.240	0.126	8.673	***	1.909

Table 5: Racial bias analysis of the pre-trained models fine-tuned on the combined human-annotated datasets (left) and combined GPT-4 (right) annotated datasets using general prompt annotation.

Class	Model	Human annotated					Llama-3 annotated (general prompt)				
		$\widehat{p_{iblack}}$	$\widehat{p_{iwhite}}$	t	p	$\frac{\widehat{p_{iblack}}}{\widehat{p_{iwhite}}}$	$\widehat{p_{iblack}}$	$\widehat{p_{iwhite}}$	t	p	$\frac{\widehat{p_{iblack}}}{\widehat{p_{iwhite}}}$
Hate	BERT	0.431	0.229	13.204	***	1.878	0.443	0.232	13.267	***	1.905
	BERTweet	0.379	0.174	14.415	***	2.185	0.371	0.169	13.59	***	2.194
	HateBERT	0.406	0.229	12.554	***	1.768	0.411	0.223	12.784	***	1.843

Table 6: Racial bias analysis of the pre-trained models fine-tuned on the combined human-annotated datasets (left) and combined Llama-3 (right) annotated datasets using general prompt annotation.

to negative classes at a higher rate than tweets written in SAE, and while some classifiers trained on LLM annotated datasets can decrease this rate, most increase it.

Bias - Subgroup & BPSN AUCs From Table 21 of the Appendix, For GMB-Subgroup-AUC, BERTweet consistently outperforms other models in six, five, and six datasets for general, FS, and CoT prompt annotations, respectively, indicating that it is most successful at accurately classifying tweets written in AAE and SAE. Additionally, BERTweet also outperforms other models on five datasets for both general and FS annotation and on all datasets annotated using CoT reasoning annotation in Generalized Mean Bias with BPSN AUC, suggesting that BERTweet is less likely to confuse non-offensive tweets written in AAE with offensive tweets not written in AAE, i.e., BERTweet significantly reduces false positive rate. The GPT-4 general prompt annotation results are consistent with the Llama-3 annotated datasets using general prompt annotation as shown in Table 22.

The dialect-wise BPSN AUC metric results for the datasets annotated using general prompt annotation are reported in Fig 3 of the Appendix. In Fig 3, we observe that tweets written in AAE seem to be biased toward having more false positives due to the low BPSN AUC scores than tweets written in SAE for all models for the Davidson (Fig 3a), Founta (Fig 3b), HatEval (Fig 3c), and AbusEval (Fig 3d) datasets. The BERTweet model is biased toward more false positives for tweets written in SAE (1%) when compared to the AAE tweets (2%) for the Golbeck (Fig 3e) and OffensEval (Fig 3f) datasets. All models are biased toward having more false positives for the SAE tweets than AAE in the Waseem (Fig 3g) dataset. The BPSN-AUC metric results for FS prompt annotation are consistent with the general prompt annotation except for the change in direction in the Waseem dataset and the increased difference in BPSN scores of the BERTweet and HateBERT models across the two groups as shown in Fig 4 of the Appendix indicating more false positives towards SAE tweets. Fig 5 in the Appendix show the BPSN AUC metric results for the CoT prompt annotation which is consistent with general and FS prompt annotation except in the Waseem (Fig 5g) where there is no large difference in the BPSN scores of the AAE and SAE tweets and in AbusEval (Fig 5d) where the BERTweet and HateBERT models are more biased towards SAE tweets than AAE tweets. In the case of Llama-3 shown in Fig 6 of the Appendix, models fine-tuned on the (Fig 6a), Fig 6b), and (Fig 6c) datasets have more false positives towards AAE tweets. There is a change in direction for models fine-tuned on the AbusEval (Fig 6d), OffensEval (Fig 6f), Golbeck (Fig 6e), and Waseem (Fig 6g) datasets, in that they have more false positives towards SAE tweets.

Overall, for GPT-4, the BERTweet model achieves an increased BPSN AUC performance across all datasets and annotation strategies, suggesting its ability to reduce false positive rate as also shown in the GMB-BPSN column in Table 21. AAE tweets seem biased toward having more false positives in most models and datasets, as shown in Fig 3. FS annotation increases dialect-wise (figures are not shown due to space) Subgroup AUC score in almost all models and datasets indicating its effect in correctly separating offensive and non-offensive AAE and SAE tweets as also seen under GMB-Subgroup-AUC summary in Table 21 when compared to general annotation. A similar observation is made for CoT annotation. FS annotation improves the ability to reduce false positives, as seen in the increment in BPSN AUC score across models, datasets, and groups (AAE and SAE) compared to general annotation as shown in Fig 4. From Fig 5, we see a similar trend for CoT annotation. For Llama-3, From Table 22, with high GMB-BPSN scores, BERTweet performs better in reducing false positives in six out of the

seven datasets. Like GPT-4, AAE tweets are biased toward having more false positives in three of the seven datasets.

5 Broader Perspectives

Our work has implications for data annotation using LLMs. As data annotation is a labor-intensive and time-consuming process for various NLP tasks, especially in online hate speech annotation where annotators can be exposed to hateful content that can negatively affect them (Vidgen et al. 2019), researchers have explored the use of LLMs such as GPT-3 to reduce the cost of annotation and have shown that downstream models trained on LLM-annotated data can achieve good performance (Wang et al. 2021). As we have demonstrated empirically in this study, if LLM-annotated datasets are biased, they may cause unfair treatment of certain groups, such as African Americans who write in African American English. If the biased data is used to train hate speech detection models used in online social platforms such as Twitter (X), tweets written in AAE might be flagged as hateful more frequently than tweets written in SAE, which could lead to the marginalization of users who use AAE in voicing their opinions or struggles for example during social movements (Tyler and Smith 1995; Thompson 2002) through the removal of such content by the platform.

6 Future Work

From the findings of this paper, we do not encourage the use of LLMs for annotating highly subjective datasets, as in the case of hate speech detection. However, they have potential, especially when looking at the classifier obtained from fine-tuning BERTweet on LLM annotated datasets. While the classifier propagates racial bias, it tends to have the best performance in terms of GMB-subgroup AUC and GMB-BPSN AUC; due to the GMB-BPSN performance, it effectively reduces false positives towards AAE tweets. This could be because BERTweet was trained on billions of (X) Twitter data as the platform contains many causal conversations and dialectal speech (Eisenstein 2017). Future work could explore fine-tuning LLMs on dialectal text; this could help reduce bias in two ways. 1) Our analysis has shown the poor annotation quality (see Table 4) of LLM annotations when compared to human annotation, which could be the source of bias. Fine-tuning LLMs on dialectal text could help improve annotation quality and potentially 2) reduce bias in downstream models. The analysis of prompting strategies and their role in introducing and mitigating bias is an open question to be answered (Torres et al. 2024). LLMs are trained using alignment training such as Reinforcement learning from human feedback (RLHF) (Ouyang et al. 2022) and Direct Preference Optimization (DPO) (Rafailov et al. 2024); future work can investigate how these training approaches can mitigate racial bias in LLMs (Zhao, Wang, and Wang 2025). Tong et al. (Tong et al. 2024) introduced the use of biased and anti-biased experts (small LMs) incorporated a debiasing signal used in bias reduction in decoding time, investigating how fine-tuning LLMs on custom datasets and the use of small LMs fine-tuned on dialectal datasets such as AAE could help in reducing racial bias (Raza, Raval, and Chatrath 2024). Finally, methods that post-process the LLM annotated data to correct for bias can be explored.

7 Limitations

Our study has several limitations. First, the subsets of the race dataset (Blodgett, Green, and O'Connor 2016) used in the training of AAEBERTweet and the fine-tuning of AAE-BERTweet to obtain the dialect classifier do not strictly contain tweets written in AAE by African-Americans and tweets written in SAE by white Americans. Also, not all African Americans use AAE, and not all AAE users are African-American, although its use suggests closeness to the African-American community (Blodgett, Green, and O'Connor 2016). Therefore, the dialect classifier can predict a tweet to be AAE even though it is SAE and vice versa. Second, our analyses are limited to the three language models fine-tuned for hate speech detection. Third, the three pre-trained models are trained on subsets of the seven hate speech detection datasets considered, and the bias assessed in the resulting classifiers could be lower bound estimates. Fourth, while we generalize our results to Llama-3, we have only focused on one annotation strategy (general) for the LLM. Fifth, due to the number of datasets studied, we settled for a prompt that worked across all datasets, which might not have been the best for each dataset for each of the prompt annotation strategies considered. Finally, the FS exemplars used in the FS and CoT annotations used the human ground-truth label. We didn't change the label. The human ground-truth label could be wrong, which might have affected the LLM's annotation; we did not make any corrections to simulate a real-world annotation task where a team or an organization utilizing FS or CoT annotation has labeled a few examples. As our evaluation is racial bias, it is essential to extend our analysis to other biases.

8 Conclusions

In this paper, we have shown that large language models, even though they are capable of annotating data for online hate detection, can introduce racial bias to the data annotation process, which can lead to unfair treatment of already marginalized groups. We used GPT-4 and Llama-3 to re-annotate seven hate speech detection datasets using general, few-shot learning, and chain-of-thought reasoning prompt annotation strategies. We then used these LLM-annotated datasets to fine-tune three pre-trained models (BERT, BERTweet and HateBERT); the analysis of resulting classifiers helped in the understanding of racial bias in LLM annotation and its propagation in the downstream models trained on them.

Ethical Statement

The datasets used in this work are all publicly available and open access. The dataset collected follows legal standards (Aliapoulios et al. 2021) and standards for ethical research (Rivers and Lewis 2014). We respect users' privacy and do not publish identifiable information about Tweet authors.

Acknowledgments

This work is partially supported by National Science Foundation (NSF) under the Grant No. 2239605, 2228616, 2228617 and 2114920.

References

Aliapoulios, M.; Bevensee, E.; Blackburn, J.; Bradlyn, B.; De Cristofaro, E.; Stringhini, G.; and Zannettou, S. 2021. A large open dataset from the Parler social network. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 15, 943–951.

Badjatiya, P.; Gupta, S.; Gupta, M.; and Varma, V. 2017. Deep learning for hate speech detection in tweets. In *Proceedings of the 26th international conference on World Wide Web companion*, 759–760.

Basile, V.; Bosco, C.; Fersini, E.; Nozza, D.; Patti, V.; Pardo, F. M. R.; Rosso, P.; and Sanguinetti, M. 2019. Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In *Proceedings of the 13th international workshop on semantic evaluation*, 54–63.

Bergsieker, H. B.; Leslie, L. M.; Constantine, V. S.; and Fiske, S. T. 2012. Stereotyping by omission: eliminate the negative, accentuate the positive. *Journal of personality and social psychology*, 102(6): 1214.

Blodgett, S. L.; Green, L.; and O'Connor, B. 2016. Demographic Dialectal Variation in Social Media: A Case Study of African-American English. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 1119–1130.

Blodgett, S. L.; and O'Connor, B. 2017. Racial disparity in natural language processing: A case study of social media african-american english. *arXiv preprint arXiv:1707.00061*.

Borkan, D.; Dixon, L.; Sorensen, J.; Thain, N.; and Vasserman, L. 2019. Nuanced metrics for measuring unintended bias with real data for text classification. In *Companion proceedings of the 2019 world wide web conference*, 491–500.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901.

Caselli, T.; Basile, V.; Mitrović, J.; and Granitzer, M. 2021. HateBERT: Retraining BERT for Abusive Language Detection in English. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, 17–25.

Caselli, T.; Basile, V.; Mitrović, J.; Kartoziya, I.; and Granitzer, M. 2020. I feel offended, don't be abusive! implicit/explicit messages in offensive and abusive language. In *Proceedings of the 12th language resources and evaluation conference*, 6193–6202.

Davidson, T.; Bhattacharya, D.; and Weber, I. 2019. Racial Bias in Hate Speech and Abusive Language Detection Datasets. In *Proceedings of the Third Workshop on Abusive Language Online*, 25–35.

Davidson, T.; Warmsley, D.; Macy, M.; and Weber, I. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the international AAAI conference on web and social media*, volume 11, 512–515.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Ding, B.; Qin, C.; Liu, L.; Chia, Y. K.; Li, B.; Joty, S.; and Bing, L. 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)*, 11173–11195.

Eisenstein, J. 2017. Identifying regional dialects in on-line social media. *The handbook of dialectology*, 368–383.

FORCE11. 2020. The FAIR Data principles. https://force11.org/info/the-fair-data-principles/. Accessed: 2025-04-01.

Fortuna, P.; and Nunes, S. 2018. A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)*, 51(4): 1–30.

Founta, A. M.; Djouvas, C.; Chatzakou, D.; Leontiadis, I.; Blackburn, J.; Stringhini, G.; Vakali, A.; Sirivianos, M.; and Kourtellis, N. 2018. Large scale crowdsourcing and characterization of twitter abusive behavior. In *Twelfth International AAAI Conference on Web and Social Media*.

Gebru, T.; Morgenstern, J.; Vecchione, B.; Vaughan, J. W.; Wallach, H.; Iii, H. D.; and Crawford, K. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12): 86–92.

Gehman, S.; Gururangan, S.; Sap, M.; Choi, Y.; and Smith, N. A. 2020. RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, 3356–3369.

Ghavami, N.; and Peplau, L. A. 2013. An intersectional analysis of gender and ethnic stereotypes: Testing three hypotheses. *Psychology of Women Quarterly*, 37(1): 113–127.

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

Golbeck, J.; Ashktorab, Z.; Banjo, R. O.; Berlinger, A.; Bhagwan, S.; Buntain, C.; Cheakalos, P.; Geller, A. A.; Gnanasekaran, R. K.; Gunasekaran, R. R.; et al. 2017. A large labeled corpus for online harassment research. In *Proceedings of the 2017 ACM on web science conference*, 229–233.

Guo, K.; Hu, A.; Mu, J.; Shi, Z.; Zhao, Z.; Vishwamitra, N.; and Hu, H. 2023. An investigation of large language models for real-world hate speech detection. In *2023 International Conference on Machine Learning and Applications (ICMLA)*, 1568–1573. IEEE.

He, X.; Zannettou, S.; Shen, Y.; and Zhang, Y. 2023. You Only Prompt Once: On the Capabilities of Prompt Learning on Large Language Models to Tackle Toxic Content. In *2024 IEEE Symposium on Security and Privacy (SP)*, 61–61. IEEE Computer Society.

Hofmann, V.; Kalluri, P. R.; Jurafsky, D.; and King, S. 2024. Dialect prejudice predicts AI decisions about people's character, employability, and criminality. *arXiv preprint arXiv:2403.00742*.

Huang, F.; Kwak, H.; and An, J. 2023. Is chatgpt better than human annotators? potential and limitations of chatgpt in explaining implicit hate speech. In *Companion proceedings of the ACM web conference 2023*, 294–297.

Kapoor, R.; Kumar, Y.; Rajput, K.; Shah, R. R.; Kumaraguru, P.; and Zimmermann, R. 2019. Mind your language: Abuse and offense detection for code-switched languages. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, 9951–9952.

Kocoń, J.; Cichecki, I.; Kaszyca, O.; Kochanek, M.; Szydło, D.; Baran, J.; Bielaniewicz, J.; Gruza, M.; Janz, A.; Kanclerz, K.; et al. 2023. ChatGPT: Jack of all trades, master of none. *Information Fusion*, 99: 101861.

Lee, Y.; Yoon, S.; and Jung, K. 2018. Comparative Studies of Detecting Abusive Language on Twitter. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, 101–106.

Li, L.; Fan, L.; Atreja, S.; and Hemphill, L. 2024. "HOT" ChatGPT: The promise of ChatGPT in detecting and discriminating hateful, offensive, and toxic comments on social media. *ACM Transactions on the Web*, 18(2): 1–36.

Li, M.; Liao, S.; Okpala, E.; Tong, M.; Costello, M.; Cheng, L.; Hu, H.; and Luo, F. 2021. COVID-HateBERT: a Pretrained Language Model for COVID-19 related Hate Speech Detection. In *2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA)*, 233–238. IEEE.

Liao, S.; Okpala, E.; Cheng, L.; Li, M.; Vishwamitra, N.; Hu, H.; Luo, F.; and Costello, M. 2023. Analysis of COVID-19 Offensive Tweets and Their Targets. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 4473–4484.

Liu, P.; Li, W.; and Zou, L. 2019. NULI at SemEval-2019 Task 6: Transfer Learning for Offensive Language Detection using Bidirectional Transformers. In *SemEval@ NAACL-HLT*, 87–91.

Maity, K.; Poornash, A.; Bhattacharya, S.; Phosit, S.; Kongsamlit, S.; Saha, S.; and Pasupa, K. 2024. HateThaiSent: Sentiment-Aided Hate Speech Detection in Thai Language. *IEEE Transactions on Computational Social Systems*.

Massey, D. S.; and Lundy, G. 2001. Use of Black English and racial discrimination in urban housing markets: New methods and findings. *Urban affairs review*, 36(4): 452–469.

Mathew, B.; Saha, P.; Yimam, S. M.; Biemann, C.; Goyal, P.; and Mukherjee, A. 2021. Hatexplain: A benchmark dataset for explainable hate speech detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, 14867–14875.

Mei, X.; Meng, C.; Liu, H.; Kong, Q.; Ko, T.; Zhao, C.; Plumbley, M. D.; Zou, Y.; and Wang, W. 2023. Wavcaps: A chatgpt-assisted weakly-labelled audio captioning dataset for audio-language multimodal research. *arXiv preprint arXiv:2303.17395*.

Meta, A. 2024. Introducing meta llama 3: The most capable openly available llm to date. *Meta AI*.

Mozafari, M.; Farahbakhsh, R.; and Crespi, N. 2020. Hate speech detection and racial bias mitigation in social media based on BERT model. *PloS one*, 15(8): e0237861.

Nguyen, D. Q.; Vu, T.; and Nguyen, A. T. 2020. BERTweet: A pre-trained language model for English Tweets. *arXiv preprint arXiv:2005.10200*.

Okpala, E.; Cheng, L.; Mbwambo, N.; and Luo, F. 2022. AAEBERT: Debiasing BERT-based Hate Speech Detection Models via Adversarial Learning. In *2022 21st IEEE International Conference on Machine Learning and Applications (ICMLA)*, 1606–1612. IEEE.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744.

Park, J. H.; Shin, J.; and Fung, P. 2018. Reducing Gender Bias in Abusive Language Detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2799–2804.

Preoţiuc-Pietro, D.; and Ungar, L. 2018. User-level race and ethnicity predictors from twitter text. In *Proceedings of the 27th international conference on computational linguistics*, 1534–1545.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I.; et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8): 9.

Rafailov, R.; Sharma, A.; Mitchell, E.; Manning, C. D.; Ermon, S.; and Finn, C. 2024. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36.

Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; and Liu, P. J. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140): 1–67.

Raza, S.; Raval, A.; and Chatrath, V. 2024. MBIAS: Mitigating Bias in Large Language Models While Retaining Context. *arXiv preprint arXiv:2405.11290*.

Rickford, J. R.; and King, S. 2016. Language and linguistics on trial: Hearing Rachel Jeantel (and other vernacular speakers) in the courtroom and beyond. *Language*, 948–988.

Rivers, C. M.; and Lewis, B. L. 2014. Ethical research standards in a world of big data. *F1000Research*, 3: 38.

Rønningstad, E.; Velldal, E.; and Øvrelid, L. 2024. A GPT among Annotators: LLM-based Entity-Level Sentiment Annotation. Association for Computational Linguistics.

Sap, M.; Card, D.; Gabriel, S.; Choi, Y.; and Smith, N. A. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, 1668–1678.

Schick, T.; Udupa, S.; and Schütze, H. 2021. Self-diagnosis and self-debiasing: A proposal for reducing corpus-based bias in nlp. *Transactions of the Association for Computational Linguistics*, 9: 1408–1424.

Schmidt, A.; and Wiegand, M. 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*, 1–10.

Schulman, J.; Zoph, B.; Kim, C.; Hilton, J.; Menick, J.; Weng, J.; Uribe, J. F. C.; Fedus, L.; Metz, L.; Pokorny, M.; et al. 2022. Chatgpt: Optimizing language models for dialogue. *OpenAI blog*, 2: 4.

Thapa, S.; Naseem, U.; and Nasim, M. 2023. From humans to machines: can chatgpt-like llms effectively replace human annotators in nlp tasks. In *Workshop Proceedings of the 17th International AAAI Conference on Web and Social Media*.

Thomas, K.; Akhawe, D.; Bailey, M.; Boneh, D.; Bursztein, E.; Consolvo, S.; Dell, N.; Durumeric, Z.; Kelley, P. G.; Kumar, D.; et al. 2021. Sok: Hate, harassment, and the changing landscape of online abuse. In *2021 IEEE Symposium on Security and Privacy (SP)*, 247–267. IEEE.

Thompson, N. 2002. Social movements, social justice and social work. *British journal of social work*, 32(6): 711–722.

Tong, S.; Zemour, E.; Lohanimit, R.; and Kagal, L. 2024. Towards Resource Efficient and Interpretable Bias Mitigation in Large Language Models. *arXiv preprint arXiv:2412.01711*.

Torres, N.; Ulloa, C.; Araya, I.; Ayala, M.; and Jara, S. 2024. A comprehensive analysis of gender, racial, and prompt-induced biases in large language models. *International Journal of Data Science and Analytics*, 1–38.

Tyler, T. R.; and Smith, H. J. 1995. Social justice and social movements.

Vaidya, A.; Mai, F.; and Ning, Y. 2020. Empirical analysis of multi-task learning for reducing identity bias in toxic comment detection. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, 683–693.

Van Aken, B.; Risch, J.; Krestel, R.; and Löser, A. 2018. Challenges for Toxic Comment Classification: An In-Depth Error Analysis. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, 33–42.

Vidgen, B.; Harris, A.; Nguyen, D.; Tromble, R.; Hale, S.; and Margetts, H. 2019. Challenges and frontiers in abusive content detection. Association for Computational Linguistics.

Vishwamitra, N.; Guo, K.; Romit, F. T.; Ondracek, I.; Cheng, L.; Zhao, Z.; and Hu, H. 2024. Moderating New Waves of Online Hate with Chain-of-Thought Reasoning in Large Language Models. In *2024 IEEE Symposium on Security and Privacy (SP)*, 178–178. IEEE Computer Society.

Wang, S.; Liu, Y.; Xu, Y.; Zhu, C.; and Zeng, M. 2021. Want To Reduce Labeling Cost? GPT-3 Can Help. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, 4195–4205.

Waseem, Z. 2016. Are you a racist or am i seeing things? annotator influence on hate speech detection on twitter. In *Proceedings of the first workshop on NLP and computational social science*, 138–142.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q. V.; Zhou, D.; et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35: 24824–24837.

Xia, M.; Field, A.; and Tsvetkov, Y. 2020. Demoting Racial Bias in Hate Speech Detection. In *Proceedings of the Eighth International Workshop on Natural Language Processing for Social Media*, 7–14.

Zampieri, M.; Malmasi, S.; Nakov, P.; Rosenthal, S.; Farra, N.; and Kumar, R. 2019. Predicting the Type and Target of Offensive Posts in Social Media. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 1415–1420.

Zhao, Y.; Wang, B.; and Wang, Y. 2025. Explicit vs. Implicit: Investigating Social Bias in Large Language Models through Self-Reflection. *arXiv preprint arXiv:2501.02295*.

Paper Checklist

1. For most authors...
   (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Yes
   (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes
   (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes
   (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? No
   (e) Did you describe the limitations of your work? Yes
   (f) Did you discuss any potential negative societal impacts of your work? Yes
   (g) Did you discuss any potential misuse of your work? No
   (h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? Yes
   (i) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes
2. Additionally, if your study involves hypotheses testing...
   (a) Did you clearly state the assumptions underlying all theoretical results? Yes
   (b) Have you provided justifications for all theoretical results? Yes

(c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? Yes

(d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? Yes, by considering alternative bias metric (AUC)

(e) Did you address potential biases or limitations in your theoretical framework? Yes in Section 7

(f) Have you related your theoretical results to the existing literature in social science? Yes

(g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? Yes

3. Additionally, if you are including theoretical proofs...

(a) Did you state the full set of assumptions of all theoretical results? NA

(b) Did you include complete proofs of all theoretical results? NA

4. Additionally, if you ran machine learning experiments...

(a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? No

(b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? Yes

(c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? No

(d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? Yes

(e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? Yes

(f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? Yes

5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, **without compromising anonymity**...

(a) If your work uses existing assets, did you cite the creators? Yes

(b) Did you mention the license of the assets? No

(c) Did you include any new assets in the supplemental material or as a URL? No

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? Yes

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? Yes

(f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR (see FORCE11 (2020))? NA

(g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset (see Gebru et al. (2021))? NA

6. Additionally, if you used crowdsourcing or conducted research with human subjects, **without compromising anonymity**...

(a) Did you include the full text of instructions given to participants and screenshots? NA

(b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? NA

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? NA

(d) Did you discuss how data is stored, shared, and deidentified? NA

A Appendix

A.1 Dialect Classifier Evaluation

We used the USERLEVELRACE dataset (Preoţiuc-Pietro and Ungar 2018) to validate our dialect classifier. The USERLEVELRACE dataset is a Twitter dataset of users who self-reported their race/ethnicity through a survey. The dataset contains 5.4M tweets from 4,132 users, of which 374 are African American (AA) and 3,184 are White. Due to X's (Twitter) terms of service, the authors could not release the actual tweets to us upon request. Instead, they provided us with each user's ID, age, gender, and race. We randomly sampled 280 AA and 280 White users. Due to the cost associated with the recent changes in X (Twitter) API, we only collected tweets from 14 AA and 14 White users within reasonable cost. For each of the users, we collect the user's most recent tweets from the user's timeline using Twarc[7], a Python library for collecting Twitter JSON data via the Twitter API[8]

Target	F1	Precision	Recall
AAE	0.611	0.832	0.483
SAE	0.746	0.636	0.903

Table 7: Performance of the dialect classification model on the dataset of users who self reported their race/ethnicity (AA and White). Evaluation metrics are macro averages.

The data collection resulted in a dataset of 14,794 tweets, of which AA users wrote 5,103 tweets and White users wrote 9,691 tweets. We sampled 5K tweets from tweets written by AA users and 5K by white users. The dialect classifier was used to predict the dialect of the sampled tweets and achieved 0.734, 0.693, and 0.679 precision, recall, and

[7]https://twarc-project.readthedocs.io/en/latest/

[8]X (Twitter)'s timeline API provides 3200 most recent tweets. Users had varying numbers of tweets; AA users tweeted 364.5 tweets on average, and White users tweeted 692.2. we could not retrieve 3200 tweets for all users, possibly due to the recent changes in X (Twitter) API or users not having enough posts.

F1 scores, respectively. The per-class performance of the dialect model on the USERLEVELRACE dataset is shown in Table 7. Upon analyzing the tweets qualitatively, we note that most AA users did not tweet in AAE, which most likely explains the low recall and F1 scores. We chose to sample 14 users from each group because, under the new X (Twitter) rules, data collection from the platform has become expensive as it is no longer free for academic research. Also, only five requests can be made per 15 minutes due to rate limitations.

Dataset	Class	Count	Sampled
Waseem	Racism	62	-
	Sexism	530	-
	Both	24	-
	Neither	5,372	500
		5,988	**1,116**
Davidson	Hate	1,430	500
	Offensive	19,190	500
	Normal	4,163	500
		24,783	**1,500**
Founta	Hateful	2,042	500
	Abusive	9,187	500
	Neither	34,329	500
		45,558	**1,500**
Golbeck	Harassment	5,285	500
	Not harassment	15,075	500
		20,360	**1,000**
OffensEval (train)	Offensive	4,640	500
	Not offensive	8,840	500
		13,240	1,000
OffensEval (test)	Offensive	240	-
	Not offensive	620	-
		860	**860**
AbusEval (train)	Explicit abuse	726	500
	Implicit abuse	2,023	500
	Not abusive	10,491	500
		13,240	**1,500**
AbusEval (test)	Explicit abuse	106	-
	Implicit abuse	72	-
	Not abusive	682	-
		860	**860**
HatEval (train)	Hate	3,783	500
	Not hate	5,217	500
		9,000	**1,000**
HatEval (test)	Hate	1,260	500
	Not hate	1,740	500
		3,000	**1,000**

Table 8: Statistics of the datasets. "-" means original tweet count was retained.

A.2 Dialect Priming Experiment

We explore the effect of dialect information in LLM annotation, (Sap et al. 2019) used dialect priming in human annotation task where annotators are instructed to consider the dialect of a tweet as a proxy for race when deciding a label for a tweet. They infer dialect using the statistical model developed by (Blodgett, Green, and O'Connor 2016). Motivated by this idea, we included dialect in each of the annotation strategies (general, FS, and CoT), used GPT-4 and Llama-3 to re-annotate our sampled datasets as discussed in Section 3.1 and inferred dialect using our dialect model discussed in Section 3.2. For the FS examples used in the FS and CoT annotation settings, we randomly sample balanced exemplars by dialect. From each class in each dataset, we randomly sampled one exemplar from tweets inferred to be AAE and tweets inferred to be SAE, making it two exemplars per class. We slightly modify the prompts we used in the setting without dialect priming as discussed in Section 3.3 as shown in Table 36. In this experiment, we only consider the Founta, Davidson, HatEval, and OffensEval datasets because they consistently showed racial bias in models fine-tuned on them as seen in the hypothesis and AUC result section 4 and 4. Due to space and the reason outlined in the prompt annotation Section 3.3, we only show the hypothesis-based results for the BERT model for the general prompt annotation in Table 32 for GPT-4 and Table 34 for Llama3. The FS and CoT prompt annotations results are shown in Table 33. Results indicate that racial disparity persists and that FS and CoT annotation with dialect priming does not help mitigate bias. Instead, they increase model bias towards AAE tweets.

Prompt strategy	Examples
General prompt	Given the tweet in triple quotes: """@USER f*k you fa*t""". Do you think the tweet is hate or offensive or normal? Only answer with one of the following: hate or offensive or normal. Do not provide an explanation for your answer.
Few-shot learning prompt	Question: is the tweet in triple quotes """Johnny Rebel - If I could be a n***** for a day: URL""" hate or offensive or normal? Only respond with hate or offensive or normal. Answer: hate. ... Question: is the tweet in triple quotes """@USER fuck you faggot""" hate or offensive or normal? Only respond with hate or offensive or normal. Answer:
Chain-of-Thought prompt	Question: Is the tweet in triple quotes """@USER @USER. LOVE ISLAM. Even the trees will say "Come kill the Jews " (and other kufr). AMEN. URL""" harassment or not harassment? Answer with only one of the following classes: harassment or not harassment. The answer should be in the last words of your final sentence after your explanation. Answer: This tweet is harassment because it calls for violence towards a group of people - the jews and other kufr, who are people who don't believe in God, making them feel unsafe. So, the answer is harassment. ... Question: Is the tweet in triple quotes """RT @USER: @USER @USER @USER @USER The Jews have caused all the world's problems since time started""" harassment or not harassment? Answer with only one of the following classes: harassment or not harassment. The answer should be in the last words of your final sentence after your explanation. Answer:

Table 9: Annotation prompt samples from the Davidson and Golbeck datasets for the three prompting strategies.

Dataset	Model	F1
Waseem	BERT	**0.356**
	BERTweet	0.340
	HateBERT	0.354
Davidson	BERT	0.800
	BERTweet	**0.823**
	HateBERT	0.775
Founta	BERT	**0.826**
	BERTweet	0.820
	HateBERT	0.824
AbusEval	BERT	0.485
	BERTweet	**0.549**
	HateBERT	0.487

Table 10: Classifier performance after fine-tuning on each Llama-3-annotated dataset with multi-class labels. Evaluation metrics are macro averages.

Dataset	Model	F1
Waseem	BERT	0.790
	BERTweet	**0.816**
	HateBERT	0.721
Davidson	BERT	0.780
	BERTweet	**0.833**
	HateBERT	0.785
Founta	BERT	0.786
	BERTweet	**0.849**
	HateBERT	0.810
Golbeck	BERT	**0.599**
	BERTweet	0.439
	HateBERT	0.514
OffenEval	BERT	0.719
	BERTweet	**0.809**
	HateBERT	0.726
AbusEval	BERT	0.692
	BERTweet	**0.792**
	HateBERT	0.735
HatEval	BERT	0.677
	BERTweet	**0.726**
	HateBERT	0.671

Table 11: Classifier performance after fine-tuning on each Llama-3-annotated dataset with binary labels. Evaluation metrics are macro averages.

Dataset	Model	F1 Score		
		Gen	FS	CoT
Waseem	BERT	0.418	0.429	0.426
	BERTweet	0.426	0.428	**0.429**
	HateBERT	**0.428**	**0.433**	0.412
Davidson	BERT	0.558	0.698	0.726
	BERTweet	**0.563**	0.753	**0.767**
	HateBERT	**0.563**	**0.763**	0.762
Founta	BERT	**0.490**	0.564	0.550
	BERTweet	**0.490**	0.605	0.539
	HateBERT	0.488	**0.624**	**0.743**
AbusEval	BERT	0.407	0.491	0.518
	BERTweet	**0.504**	**0.576**	**0.647**
	HateBERT	0.471	0.529	0.577

Table 12: Classifier performance after fine-tuning on each GPT-annotated dataset with multi-class labels for each prompting strategy. Evaluation metrics are macro averages.

Dataset	Model	F1
Waseem	BERT	0.409
	BERTweet	**0.423**
	HateBERT	0.416
Davidson	BERT	0.693
	BERTweet	0.771
	HateBERT	**0.783**
Founta	BERT	**0.738**
	BERTweet	0.713
	HateBERT	0.716
AbusEval	BERT	0.457
	BERTweet	**0.570**
	HateBERT	0.518

Table 13: Classifier performance after fine-tuning on each human-annotated dataset with multi-class labels. Evaluation metrics are macro averages.

Dataset	Model	F1
Waseem	BERT	0.838
	BERTweet	**0.852**
	HateBERT	**0.852**
Davidson	BERT	0.838
	BERTweet	**0.888**
	HateBERT	0.885
Founta	BERT	0.859
	BERTweet	**0.865**
	HateBERT	0.861
Golbeck	BERT	0.527
	BERTweet	0.575
	HateBERT	**0.601**
OffenEval	BERT	0.636
	BERTweet	**0.715**
	HateBERT	0.657
AbusEval	BERT	0.635
	BERTweet	**0.748**
	HateBERT	0.657
HatEval	BERT	0.503
	BERTweet	**0.527**
	HateBERT	0.449

Table 14: Classifier performance after fine-tuning on each human-annotated dataset with binary labels. Evaluation metrics are macro averages.

Dataset	Precision	Recall	F1	Accuracy
Davidson	0.75	0.61	0.55	0.61
Founta	0.60	0.46	0.38	0.46
AbusEval	0.60	0.55	0.55	0.61
Waseem	0.80	0.63	0.68	0.79
Golbeck	0.65	0.65	0.64	0.65
OffensEval	0.76	0.73	0.73	0.76
HatEval	0.75	0.75	0.75	0.75

Table 15: Performance of human vs GPT-4 general prompt annotation. Evaluation metrics are macro averages. In all of the metrics, the level of agreement is above 50% except in the recall, F1 and accuracy metrics of the Founta dataset.

	AAE				SAE			
Dataset	Precision	Recall	F1	Accuracy	Precision	Recall	F1	Accuracy
Davidson	0.70	0.55	0.50	0.62	0.73	0.63	0.54	0.61
Founta	0.55	0.48	0.34	0.35	0.60	0.45	0.38	0.48
AbusEval	0.50	0.50	0.50	0.65	0.61	0.56	0.55	0.60
Waseem	0.78	0.68	0.71	0.79	0.80	0.61	0.67	0.79
Golbeck	0.68	0.68	0.68	0.68	0.65	0.64	0.64	0.64
OffensEval	0.82	0.81	0.81	0.82	0.75	0.72	0.72	0.75
HatEval	0.74	0.72	0.73	0.75	0.75	0.75	0.75	0.75

Table 16: Performance of human vs GPT-4 general prompt annotation when datasets are conditioned on dialect. Evaluation metrics are macro averages. In general, the rate at which GPT-4 general prompt annotation aligns with human-annotation is slightly better if tweets are written in SAE in terms of precision, F1, and accuracy for most datasets.

Dataset	Precision	Recall	F1	Accuracy
Davidson	0.78	0.75	0.75	0.75
Founta	0.64	0.51	0.46	0.51
AbusEval	0.58	0.57	0.57	0.64
Waseem	0.80	0.75	0.78	0.84
Golbeck	0.67	0.67	0.67	0.67
OffensEval	0.78	0.75	0.76	0.78
HatEval	0.76	0.76	0.76	0.76

Table 17: Performance of human vs GPT-4 few-shot learning prompt annotation. Evaluation metrics are macro averages.

	AAE				SAE			
Dataset	Precision	Recall	F1	Accuracy	Precision	Recall	F1	Accuracy
Davidson	0.71	0.67	0.68	0.71	0.77	0.77	0.75	0.78
Founta	0.59	0.61	0.52	0.54	0.65	0.48	0.43	0.50
AbusEval	0.50	0.52	0.50	0.68	0.59	0.57	0.58	0.64
Waseem	0.83	0.88	0.82	0.87	0.80	0.73	0.76	0.84
Golbeck	0.71	0.71	0.71	0.71	0.67	0.67	0.67	0.67
OffensEval	0.86	0.86	0.86	0.86	0.77	0.74	0.74	0.77
HatEval	0.75	0.74	0.75	0.76	0.76	0.76	0.76	0.76

Table 18: Performance of human vs GPT-4 few-shot learning prompt annotation when datasets are conditioned on dialect. Evaluation metrics are macro averages.

Dataset	Precision	Recall	F1	Accuracy
Davidson	0.79	0.76	0.76	0.76
Founta	0.69	0.60	0.56	0.60
AbusEval	0.62	0.61	0.61	0.66
Waseem	0.81	0.74	0.77	0.84
Golbeck	0.67	0.66	0.66	0.66
OffensEval	0.79	0.71	0.72	0.76
HatEval	0.79	0.78	0.78	0.78

Table 19: Performance of human vs GPT-4 chain-of-thought prompt annotation. Evaluation metrics are macro averages.

	AAE				SAE			
Dataset	Precision	Recall	F1	Accuracy	Precision	Recall	F1	Accuracy
Davidson	0.77	0.71	0.73	0.74	0.78	0.77	0.75	0.77
Founta	0.69	0.74	0.67	0.68	0.68	0.56	0.51	0.58
AbusEval	0.54	0.55	0.53	0.70	0.62	0.61	0.61	0.66
Waseem	0.86	0.88	0.85	0.87	0.81	0.71	0.75	0.84
Golbeck	0.69	0.69	0.69	0.69	0.66	0.66	0.66	0.66
OffensEval	0.88	0.87	0.87	0.88	0.78	0.69	0.69	0.74
HatEval	0.81	0.80	0.80	0.81	0.78	0.78	0.77	0.77

Table 20: Performance of human vs GPT-4 chain-of-thought prompt annotation when datasets are conditioned on dialect. Evaluation metrics are macro averages.

		F1 Score			GMB-Sub			GMB-BPSN		
Dataset	Model	Gen	FS	CoT	Gen	FS	CoT	Gen	FS	CoT
Waseem	BERT	0.855	0.888	0.866	0.916	0.957	0.960	0.888	0.953	0.960
	BERTweet	0.876	**0.901**	**0.873**	**0.930**	**0.974**	**0.912**	0.892	**0.977**	**0.967**
	HateBERT	**0.880**	0.892	0.867	0.909	0.948	0.823	**0.932**	0.956	0.962
Davidson	BERT	0.823	0.841	0.849	0.808	0.856	0.868	0.724	0.739	0.798
	BERTweet	**0.857**	**0.900**	**0.900**	**0.926**	**0.937**	0.935	**0.892**	**0.916**	**0.900**
	HateBERT	0.840	0.855	0.864	0.922	0.924	**0.940**	0.840	0.857	0.868
Founta	BERT	0.686	0.736	0.857	0.830	0.859	0.908	0.737	0.786	0.875
	BERTweet	**0.756**	**0.811**	**0.865**	0.823	0.872	**0.914**	0.790	0.828	**0.901**
	HateBERT	0.743	0.772	0.834	**0.841**	**0.901**	0.910	**0.813**	**0.874**	0.879
Golbeck	BERT	**0.692**	**0.737**	**0.654**	0.798	**0.832**	0.780	0.783	**0.852**	0.789
	BERTweet	0.576	0.648	0.643	**0.802**	0.822	**0.811**	**0.788**	0.827	**0.820**
	HateBERT	0.633	0.686	0.601	0.758	0.770	0.771	0.741	0.789	0.772
OffenEval	BERT	0.605	0.642	0.613	0.775	0.838	0.806	0.771	0.841	0.818
	BERTweet	**0.732**	**0.736**	**0.662**	**0.815**	**0.874**	**0.912**	**0.819**	**0.868**	**0.900**
	HateBERT	0.625	0.630	0.578	0.787	0.815	0.823	0.787	0.812	0.800
AbusEval	BERT	0.669	0.700	0.695	0.770	0.807	0.799	0.762	0.806	0.797
	BERTweet	**0.799**	**0.780**	**0.783**	**0.866**	**0.876**	**0.841**	**0.865**	**0.873**	**0.844**
	HateBERT	0.729	0.729	0.742	0.823	0.856	0.836	0.820	0.857	0.836
HatEval	BERT	0.698	0.674	0.639	0.758	0.730	0.736	0.587	0.554	0.671
	BERTweet	**0.773**	**0.733**	**0.766**	**0.880**	**0.853**	**0.855**	**0.856**	**0.809**	**0.849**
	HateBERT	0.715	0.684	0.644	0.787	0.783	0.773	0.600	0.592	0.703

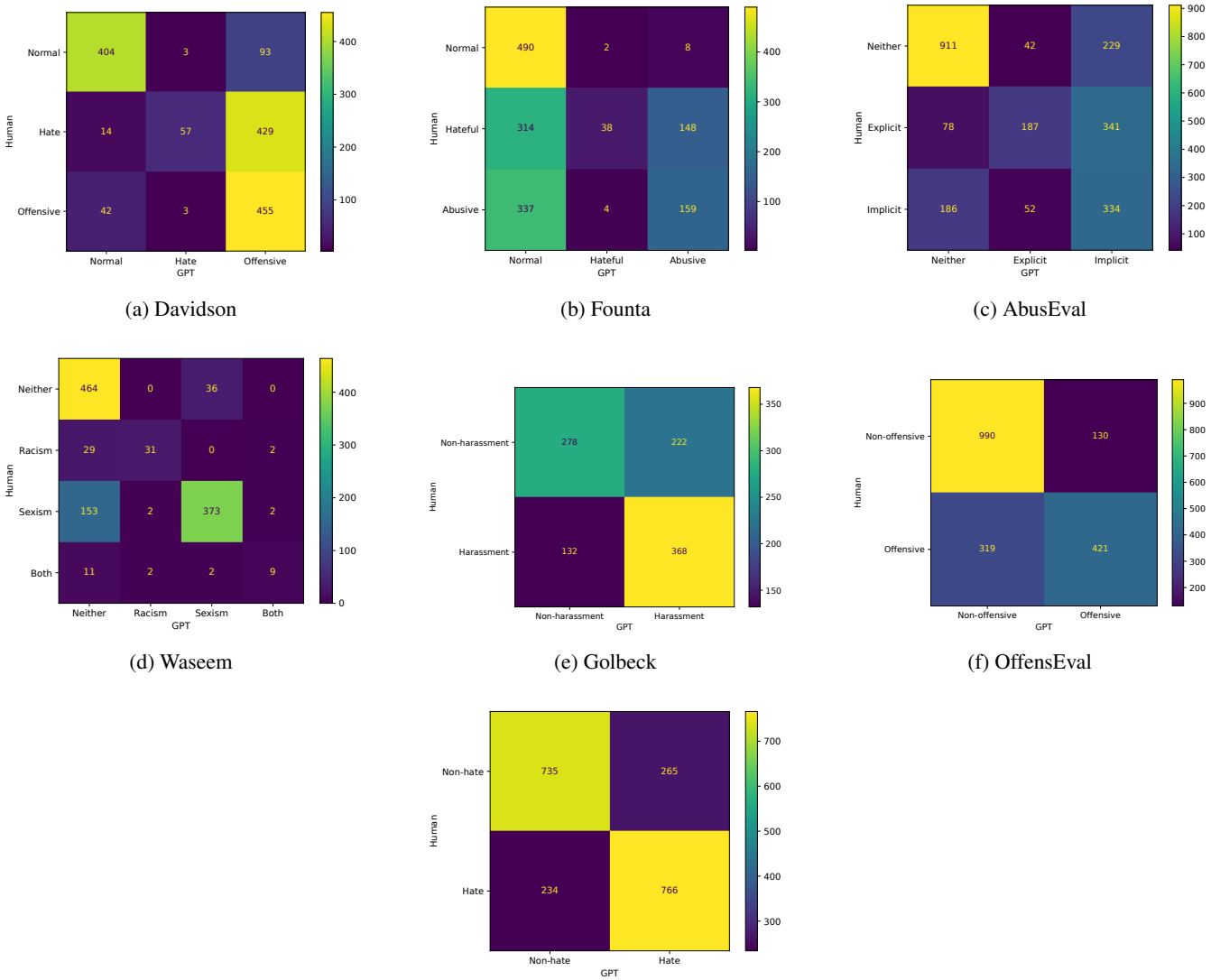
Table 21: Classifier performance after fine-tuning on each GPT-4 annotated dataset with binary labels for each prompting strategy. Evaluation metrics are macro averages.

Dataset	Model	GMB-Sub	GMB-BPSN
Waseem	BERT	0.876	0.860
	BERTweet	**0.911**	**0.898**
	HateBERT	0.900	0.836
Davidson	BERT	0.673	0.535
	BERTweet	**0.873**	**0.786**
	HateBERT	0.836	0.693
Founta	BERT	0.838	0.805
	BERTweet	**0.924**	**0.916**
	HateBERT	0.857	0.833
Golbeck	BERT	**0.766**	**0.781**
	BERTweet	0.631	0.603
	HateBERT	0.679	0.664
OffenEval	BERT	0.815	0.818
	BERTweet	**0.858**	**0.842**
	HateBERT	0.766	0.743
AbusEval	BERT	0.811	0.813
	BERTweet	**0.851**	**0.856**
	HateBERT	0.818	0.811
HatEval	BERT	0.756	0.590
	BERTweet	**0.837**	**0.807**
	HateBERT	0.752	0.569

Table 22: AUC results after fine-tuning on each Llama-3 annotated dataset with binary labels for each prompting strategy. Evaluation metrics are macro averages.

		Human annotated					GPT-4 annotated (general prompt)				
Dataset	class	$\widehat{p_{iblack}}$	$\widehat{p_{iwhite}}$	t	p	$\frac{\widehat{p_{iblack}}}{\widehat{p_{iwhite}}}$	$\widehat{p_{iblack}}$	$\widehat{p_{iwhite}}$	t	p	$\frac{\widehat{p_{iblack}}}{\widehat{p_{iwhite}}}$
AbusEval	Explicit	0.334	0.292	8	***	1.144	0.206	0.184	8.546	***	1.116
	Implicit	0.234	0.276	-7.751	***	0.848	0.285	0.327	-5.97	***	0.872
OffensEval	Offensive	0.488	0.408	8.123	***	1.195	0.386	0.275	12.007	***	1.406
HatEval	Hate	0.497	0.379	12.541	***	1.311	0.444	0.264	17.17	***	1.678
Davidson	Hate	0.345	0.331	3.173	0.002	1.042	0.050	0.051	-1.143		0.985
	Offensive	0.397	0.241	19.505	***	1.644	0.497	0.291	15.226	***	1.708
Founta	Hate	0.372	0.367	0.781		1.013	0.031	0.027	4.429	***	1.149
	Abuse	0.366	0.273	9.72	***	1.342	0.163	0.099	9.488	***	1.657
Waseem	Racism	0.099	0.104	-9.205	***	0.954	0.048	0.048	-0.905		0.992
	Sexism	0.269	0.232	5.99	***	1.158	0.124	0.112	3.432	***	1.106
	R & S	0.045	0.046	-2.483	0.013	0.981	0.025	0.025	2.476	0.013	1.036
Golbeck	Harassment	0.501	0.461	8.578	***	1.088	0.584	0.499	11.141	***	1.170

Table 23: BERT classifiers. Racial bias analysis of the pre-trained BERT model fine-tuned on human-annotated datasets (left) and GPT-4 (right) annotated datasets using general prompt annotation.



(a) Davidson (b) Founta (c) AbusEval

(d) Waseem (e) Golbeck (f) OffensEval

(g) HatEval

Figure 2: The confusion matrix of human annotation and GPT-4 annotation on full (training and testing) datasets. GPT-4 tends to label most of the tweets annotated as hate and abusive as offensive (in 29% of the total tweets) in the Davidson datasets. In the Founta dataset, GPT-4 labels a sizable number of tweets annotated as Hateful and Abusive by humans as Normal in 21% and 22% of the total tweets, respectively. GPT-4 labels 14% of the tweets as implicit abuse in the AbusEval dataset and does a good job in correctly annotating sexist tweets in the Waseem dataset. In the binary datasets, GPT-4 disagrees with humans by re-labeling harassment tweets to non-harassment in 13% of the tweets in the Golbeck dataset, re-labeling offensive as non-offensive in 17% of the tweets in the OffensEval dataset and re-labeling hate to non-hate in 12% of the tweets in the HatEval dataset.

Dataset	class	Human annotated					Llama-3 annotated (general prompt)				
		$\widehat{p_{iblack}}$	$\widehat{p_{iwhite}}$	t	p	$\frac{\widehat{p_{iblack}}}{\widehat{p_{iwhite}}}$	$\widehat{p_{iblack}}$	$\widehat{p_{iwhite}}$	t	p	$\frac{\widehat{p_{iblack}}}{\widehat{p_{iwhite}}}$
AbusEval	Explicit	0.334	0.292	8	***	1.144	0.185	0.171	4.298	***	1.082
	Implicit	0.234	0.276	-7.751	***	0.848	0.321	0.338	-2.65	0.008	0.949
OffensEval	Offensive	0.488	0.408	8.123	***	1.195	0.536	0.424	12.32	***	1.263
HatEval	Hate	0.497	0.379	12.541	***	1.311	0.461	0.283	17.726	***	1.678
Davidson	Hate	0.345	0.331	3.173	0.002	1.042	0.023	0.031	-11.502	***	0.738
	Offensive	0.397	0.241	19.505	***	1.644	0.594	0.354	17.001	***	1.679
Founta	Hate	0.372	0.367	0.781		1.013	0.026	0.031	-5.811	***	0.863
	Abuse	0.366	0.273	9.72	***	1.342	0.381	0.226	12.718	***	1.685
Waseem	Racism	0.099	0.104	-9.205	***	0.954	0.064	0.072	-7.627	***	0.889
	Sexism	0.269	0.232	5.99	***	1.158	0.613	0.530	11.661	***	1.155
	R & S	0.045	0.046	-2.483	0.013	0.981	0.059	0.049	9.426	***	1.212
Golbeck	Harassment	0.501	0.461	8.578	***	1.088	0.587	0.538	9.459	***	1.091

Table 24: BERT classifiers. Racial bias analysis of the pre-trained BERT model fine-tuned on human-annotated datasets (left) and Llama-3 (right) annotated datasets using general prompt annotation.

Dataset	class	Human annotated					GPT-4 annotated (general prompt)				
		$\widehat{p_{iblack}}$	$\widehat{p_{iwhite}}$	t	p	$\frac{\widehat{p_{iblack}}}{\widehat{p_{iwhite}}}$	$\widehat{p_{iblack}}$	$\widehat{p_{iwhite}}$	t	p	$\frac{\widehat{p_{iblack}}}{\widehat{p_{iwhite}}}$
AbusEval	Explicit	0.214	0.188	5.946	***	1.136	0.165	0.146	9.272	***	1.127
	Implicit	0.257	0.269	-5.267	***	0.957	0.231	0.221	3.032	0.002	1.044
OffensEval	Offensive	0.415	0.393	6.075	***	1.058	0.294	0.257	9.879	***	1.144
HatEval	Hate	0.426	0.331	18.85	***	1.286	0.319	0.238	13.578	***	1.343
Davidson	Hate	0.246	0.210	7.955	***	1.169	0.064	0.065	-2.179	0.029	0.988
	Offensive	0.366	0.237	22.687	***	1.544	0.568	0.427	14.721	***	1.332
Founta	Hate	0.315	0.234	16.865	***	1.343	0.049	0.038	11.38	***	1.300
	Abuse	0.459	0.441	2.146	0.032	1.042	0.217	0.141	11.993	***	1.543
Waseem	Racism	0.071	0.069	10.46	***	1.036	0.051	0.049	8.747	***	1.031
	Sexism	0.237	0.212	6.087	***	1.118	0.137	0.131	2.91	0.004	1.045
	R & S	0.052	0.051	4.854	***	1.017	0.042	0.041	6.746	***	1.025
Golbeck	Harassment	0.474	0.471	2.11	0.035	1.005	0.574	0.555	7.816	***	1.034

Table 25: BERTweet classifiers. Racial bias analysis of the pre-trained BERTweet model fine-tuned on human-annotated datasets (left) and GPT-4 (right) annotated datasets using general prompt annotation.

Dataset	class	Human annotated					Llama-3 annotated (general prompt)				
		$\widehat{p_{iblack}}$	$\widehat{p_{iwhite}}$	t	p	$\frac{\widehat{p_{iblack}}}{\widehat{p_{iwhite}}}$	$\widehat{p_{iblack}}$	$\widehat{p_{iwhite}}$	t	p	$\frac{\widehat{p_{iblack}}}{\widehat{p_{iwhite}}}$
AbusEval	Explicit	0.214	0.188	5.946	***	1.136	0.122	0.112	5.975	***	1.092
	Implicit	0.257	0.269	-5.267	***	0.957	0.286	0.275	3.414	***	1.038
OffensEval	Offensive	0.415	0.393	6.075	***	1.058	0.443	0.407	8.587	***	1.087
HatEval	Hate	0.426	0.331	18.85	***	1.286	0.399	0.298	15.398	***	1.342
Davidson	Hate	0.246	0.210	7.955	***	1.169	0.035	0.041	-11.916	***	0.867
	Offensive	0.366	0.237	22.687	***	1.544	0.619	0.467	15.187	***	1.326
Founta	Hate	0.315	0.234	16.865	***	1.343	0.037	0.035	8.297	***	1.056
	Abuse	0.459	0.441	2.146	0.032	1.042	0.440	0.285	15.158	***	1.547
Waseem	Racism	0.071	0.069	10.46	***	1.036	0.044	0.044	-1.738		0.988
	Sexism	0.237	0.212	6.087	***	1.118	0.568	0.540	5.043	***	1.052
	R & S	0.052	0.051	4.854	***	1.017	0.056	0.056	-0.17		0.999
Golbeck	Harassment	0.474	0.471	2.11	0.035	1.005	0.584	0.573	7.059	***	1.020

Table 26: BERTweet classifiers. Racial bias analysis of the pre-trained BERTweet model fine-tuned on human-annotated datasets (left) and Llama-3 (right) annotated datasets using general prompt annotation.

Dataset	class	Human annotated					GPT-4 annotated (general prompt)				
		$\widehat{p_{iblack}}$	$\widehat{p_{iwhite}}$	t	p	$\frac{\widehat{p_{iblack}}}{\widehat{p_{iwhite}}}$	$\widehat{p_{iblack}}$	$\widehat{p_{iwhite}}$	t	p	$\frac{\widehat{p_{iblack}}}{\widehat{p_{iwhite}}}$
AbusEval	Explicit	0.242	0.217	4.988	***	1.115	0.169	0.147	7.207	***	1.149
	Implicit	0.197	0.190	1.992	0.047	1.032	0.236	0.215	4.42	***	1.096
OffensEval	Offensive	0.449	0.408	7.439	***	1.100	0.344	0.281	10.661	***	1.226
HatEval	Hate	0.524	0.435	14.574	***	1.203	0.413	0.311	15.015	***	1.329
Davidson	Hate	0.228	0.180	9.036	***	1.267	0.051	0.060	-9.956	***	0.853
	Offensive	0.344	0.191	20.355	***	1.796	0.535	0.323	19.051	***	1.657
Founta	Hate	0.321	0.281	8.587	***	1.143	0.041	0.036	5.574	***	1.144
	Abuse	0.389	0.313	10.512	***	1.241	0.179	0.121	9.659	***	1.478
Waseem	Racism	0.085	0.077	7.717	***	1.107	0.058	0.052	8.613	***	1.125
	Sexism	0.265	0.236	4.858	***	1.123	0.109	0.111	-0.811		0.979
	R & S	0.051	0.052	-0.752		0.982	0.033	0.035	-2.206	0.028	0.938
Golbeck	Harassment	0.466	0.453	5.04	***	1.029	0.573	0.542	7.906	***	1.059

Table 27: HateBERT classifiers. Racial bias analysis of the pre-trained HateBERT model fine-tuned on human-annotated datasets (left) and GPT-4 (right) annotated datasets using general prompt annotation.

Dataset	class	Human annotated					Llama-3 annotated (general prompt)				
		$\widehat{p_{iblack}}$	$\widehat{p_{iwhite}}$	t	p	$\frac{\widehat{p_{iblack}}}{\widehat{p_{iwhite}}}$	$\widehat{p_{iblack}}$	$\widehat{p_{iwhite}}$	t	p	$\frac{\widehat{p_{iblack}}}{\widehat{p_{iwhite}}}$
AbusEval	Explicit	0.242	0.217	4.988	***	1.115	0.133	0.130	1.058		1.021
	Implicit	0.197	0.190	1.992	0.047	1.032	0.252	0.231	4.244	***	1.089
OffensEval	Offensive	0.449	0.408	7.439	***	1.100	0.511	0.464	8.303	***	1.102
HatEval	Hate	0.524	0.435	14.574	***	1.203	0.480	0.373	15.238	***	1.284
Davidson	Hate	0.228	0.180	9.036	***	1.267	0.029	0.040	-14.658	***	0.726
	Offensive	0.344	0.191	20.355	***	1.796	0.608	0.378	19.855	***	1.610
Founta	Hate	0.321	0.281	8.587	***	1.143	0.032	0.034	-3.472	***	0.940
	Abuse	0.389	0.313	10.512	***	1.241	0.410	0.261	14.98	***	1.568
Waseem	Racism	0.085	0.077	7.717	***	1.107	0.045	0.048	-3.59	***	0.942
	Sexism	0.265	0.236	4.858	***	1.123	0.590	0.526	7.747	***	1.120
	R & S	0.051	0.052	-0.752		0.982	0.063	0.062	0.486		1.010
Golbeck	Harassment	0.466	0.453	5.04	***	1.029	0.539	0.521	5.469	***	1.035

Table 28: HateBERT classifiers. Racial bias analysis of the pre-trained HateBERT model fine-tuned on human-annotated datasets (left) and Llama-3 (right) annotated datasets using general prompt annotation.

Dataset	class	Few-shot Prompt annotation					CoT Prompt Annotation				
		$\widehat{p_{iblack}}$	$\widehat{p_{iwhite}}$	t	p	$\frac{\widehat{p_{iblack}}}{\widehat{p_{iwhite}}}$	$\widehat{p_{iblack}}$	$\widehat{p_{iwhite}}$	t	p	$\frac{\widehat{p_{iblack}}}{\widehat{p_{iwhite}}}$
AbusEval	Explicit	0.323	0.307	3.735	***	1.053	0.325	0.310	3.806	***	1.049
	Implicit	0.157	0.164	-1.876		0.958	0.166	0.184	-4.607	***	0.903
OffensEval	Offensive	0.387	0.283	10.649	***	1.368	0.348	0.262	9.71	***	1.327
HatEval	Hate	0.439	0.277	16.322	***	1.588	0.489	0.320	16.233	***	1.529
Davidson	Hate	0.193	0.223	-10.378	***	0.865	0.189	0.207	-6.559	***	0.913
	Offensive	0.512	0.314	20.523	***	1.632	0.522	0.329	19.096	***	1.589
Founta	Hate	0.100	0.047	14.582	***	2.121	0.101	0.072	12.428	***	1.409
	Abuse	0.150	0.108	8.618	***	1.394	0.239	0.129	12.207	***	1.854
Waseem	Racism	0.078	0.081	-7.814	***	0.965	0.061	0.062	-2.165	0.031	0.983
	Sexism	0.185	0.166	5.879	***	1.119	0.170	0.138	7.861	***	1.235
	R & S	0.036	0.035	2.987	0.003	1.020	0.033	0.031	5.072	***	1.074
Golbeck	Harassment	0.552	0.456	12.274	***	1.211	0.613	0.522	13.114	***	1.175

Table 29: BERT classifiers. Racial bias analysis of the pre-trained BERT model fine-tuned on GPT-4 annotated datasets using few-shot prompt (left) and chain-of-thought prompt annotation (right).

Dataset	class	Few-shot Prompt annotation					CoT Prompt Annotation				
		$\widehat{p_{iblack}}$	$\widehat{p_{iwhite}}$	t	p	$\frac{\widehat{p_{iblack}}}{\widehat{p_{iwhite}}}$	$\widehat{p_{iblack}}$	$\widehat{p_{iwhite}}$	t	p	$\frac{\widehat{p_{iblack}}}{\widehat{p_{iwhite}}}$
AbusEval	Explicit	0.190	0.150	6.512	***	1.264	0.171	0.134	6.04	***	1.272
	Implicit	0.169	0.161	4.551	***	1.047	0.171	0.169	1.562		1.015
OffensEval	Offensive	0.316	0.270	8.715	***	1.171	0.289	0.232	12.054	***	1.246
HatEval	Hate	0.331	0.254	13.473	***	1.303	0.330	0.246	17.084	***	1.345
Davidson	Hate	0.141	0.135	2.213	0.027	1.049	0.144	0.126	6.261	***	1.147
	Offensive	0.475	0.374	19.107	***	1.269	0.483	0.377	16.802	***	1.281
Founta	Hate	0.174	0.065	24.99	***	2.693	0.169	0.076	28.812	***	2.233
	Abuse	0.173	0.096	20.796	***	1.815	0.325	0.216	15.339	***	1.505
Waseem	Racism	0.063	0.061	10.357	***	1.036	0.060	0.058	11.948	***	1.036
	Sexism	0.181	0.168	4.858	***	1.078	0.177	0.164	5.638	***	1.081
	R & S	0.050	0.048	9.919	***	1.036	0.047	0.046	9.344	***	1.028
Golbeck	Harassment	0.546	0.533	5.177	***	1.025	0.571	0.555	7.212	***	1.028

Table 30: BERTweet classifiers. Racial bias analysis of the pre-trained BERTweet model fine-tuned on GPT-4 annotated datasets using few-shot prompt (left) and chain-of-thought prompt annotation (right).

Dataset	class	Few-shot Prompt annotation					CoT Prompt Annotation				
		$\widehat{p_{iblack}}$	$\widehat{p_{iwhite}}$	t	p	$\frac{\widehat{p_{iblack}}}{\widehat{p_{iwhite}}}$	$\widehat{p_{iblack}}$	$\widehat{p_{iwhite}}$	t	p	$\frac{\widehat{p_{iblack}}}{\widehat{p_{iwhite}}}$
AbusEval	Explicit	0.217	0.179	6.527	***	1.211	0.207	0.174	5.816	***	1.192
	Implicit	0.144	0.133	4.988	***	1.083	0.153	0.146	2.815	0.005	1.045
OffensEval	Offensive	0.357	0.294	10.195	***	1.212	0.349	0.267	13.016	***	1.310
HatEval	Hate	0.419	0.323	14.512	***	1.298	0.433	0.350	14.201	***	1.239
Davidson	Hate	0.122	0.124	-0.53		0.985	0.115	0.104	2.933	0.003	1.102
	Offensive	0.478	0.283	22.179	***	1.685	0.492	0.305	20.111	***	1.615
Founta	Hate	0.112	0.055	13.134	***	2.039	0.130	0.064	14.379	***	2.029
	Abuse	0.141	0.096	10.735	***	1.475	0.276	0.192	11.634	***	1.434
Waseem	Racism	0.078	0.068	9.471	***	1.143	0.074	0.063	11.93	***	1.169
	Sexism	0.170	0.152	4.151	***	1.115	0.153	0.136	5.226	***	1.125
	R & S	0.044	0.045	-0.054		0.999	0.040	0.041	-0.152		0.996
Golbeck	Harassment	0.551	0.515	8.364	***	1.070	0.576	0.541	8.741	***	1.064

Table 31: HateBERT classifiers. Racial bias analysis of pre-trained HateBERT model fine-tuned on GPT-4 annotated datasets using few-shot prompt (left) and chain-of-thought prompt annotation (right).

Dataset	class	$\widehat{p_{iblack}}$	$\widehat{p_{iwhite}}$	t	p	$\frac{\widehat{p_{iblack}}}{\widehat{p_{iwhite}}}$
OffensEval	Offensive	0.375	0.260	11.794	***	1.444
HatEval	Hate	0.374	0.260	13.459	***	1.440
Davidson	Hate	0.070	0.075	-5.332	***	0.933
	Offensive	0.412	0.297	10.607	***	1.385
Founta	Hate	0.026	0.025	1.251		1.047
	Abuse	0.114	0.090	4.321	***	1.257

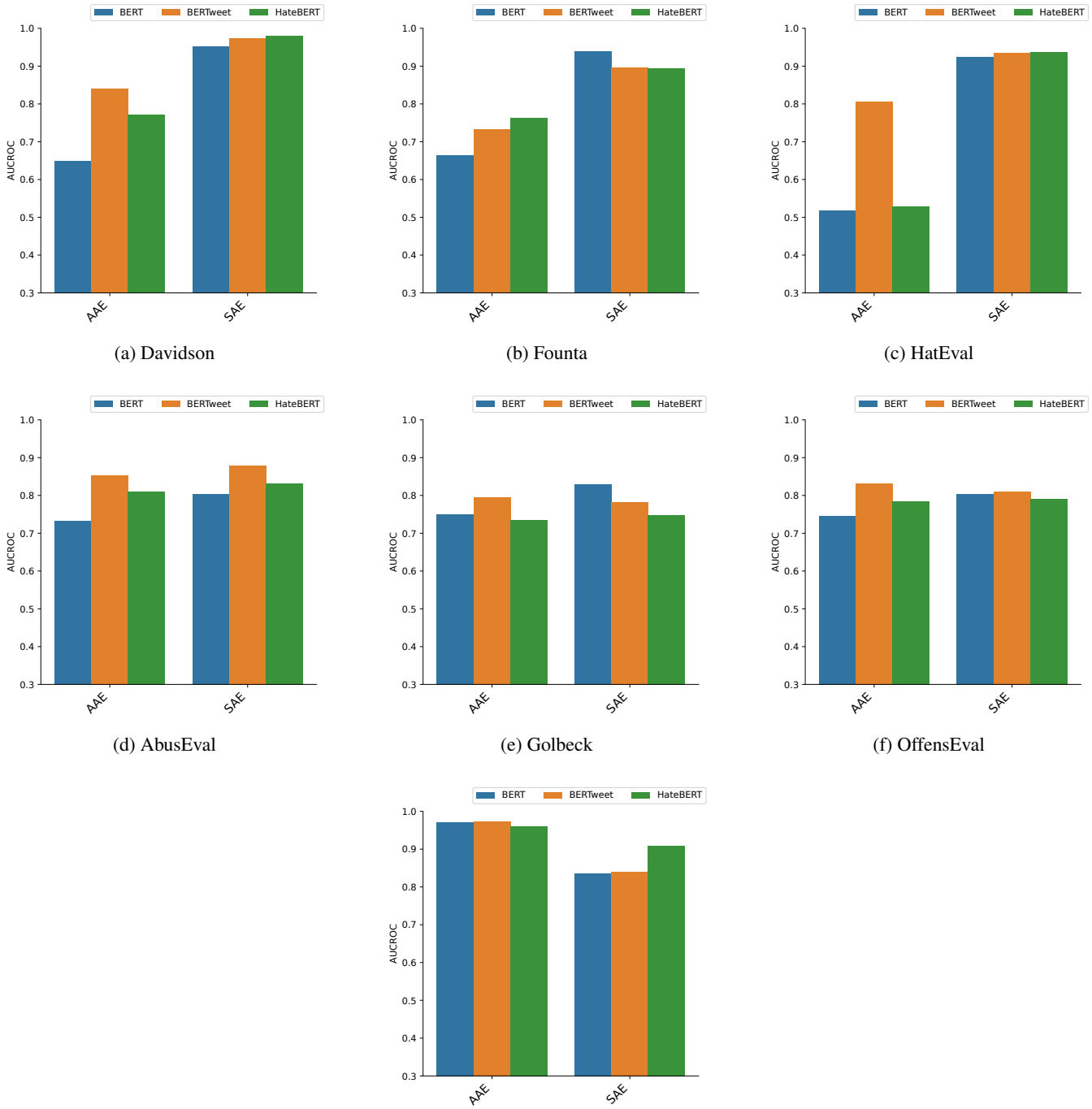
Table 32: BERT classifiers. Racial bias analysis of the pre-trained BERT model fine-tuned on GPT-4 annotated datasets using general prompt annotation with dialect priming.

		Few-shot Prompt annotation					CoT Prompt Annotation				
Dataset	class	$\widehat{p_{iblack}}$	$\widehat{p_{iwhite}}$	t	p	$\frac{\widehat{p_{iblack}}}{\widehat{p_{iwhite}}}$	$\widehat{p_{iblack}}$	$\widehat{p_{iwhite}}$	t	p	$\frac{\widehat{p_{iblack}}}{\widehat{p_{iwhite}}}$
OffensEval	Offensive	0.334	0.214	12.773	***	1.564	0.304	0.228	9.044	***	1.333
HatEval	Hate	0.388	0.212	17.361	***	1.831	0.472	0.286	17.099	***	1.652
Davidson	Hate	0.157	0.153	2.116	0.034	1.026	0.101	0.089	8.251	***	1.138
	Offensive	0.394	0.265	15.033	***	1.488	0.477	0.306	15.207	***	1.559
Founta	Hate	0.056	0.045	5.383	***	1.243	0.069	0.057	10.221	***	1.209
	Abuse	0.130	0.095	6.206	***	1.365	0.319	0.170	12.794	***	1.880

Table 33: BERT classifiers. Racial bias analysis of the pre-trained BERT model fine-tuned on GPT-4 annotated datasets using few-shot prompt (left) and chain-of-thought prompt annotation (right) with dialect priming.

Dataset	class	$\widehat{p_{iblack}}$	$\widehat{p_{iwhite}}$	t	p	$\frac{\widehat{p_{iblack}}}{\widehat{p_{iwhite}}}$
OffensEval	Offensive	0.621	0.536	11.947	***	1.157
HatEval	Hate	0.452	0.392	6.579	***	1.154
Davidson	Hate	0.024	0.029	-8.601	***	0.808
	Offensive	0.500	0.350	11.249	***	1.430
Founta	Hate	0.024	0.027	-6.193	***	0.866
	Abuse	0.406	0.285	9.108	***	1.422

Table 34: BERT classifiers. Racial bias analysis of the pre-trained BERT model fine-tuned on Llama-3 annotated datasets using general prompt annotation with dialect priming.

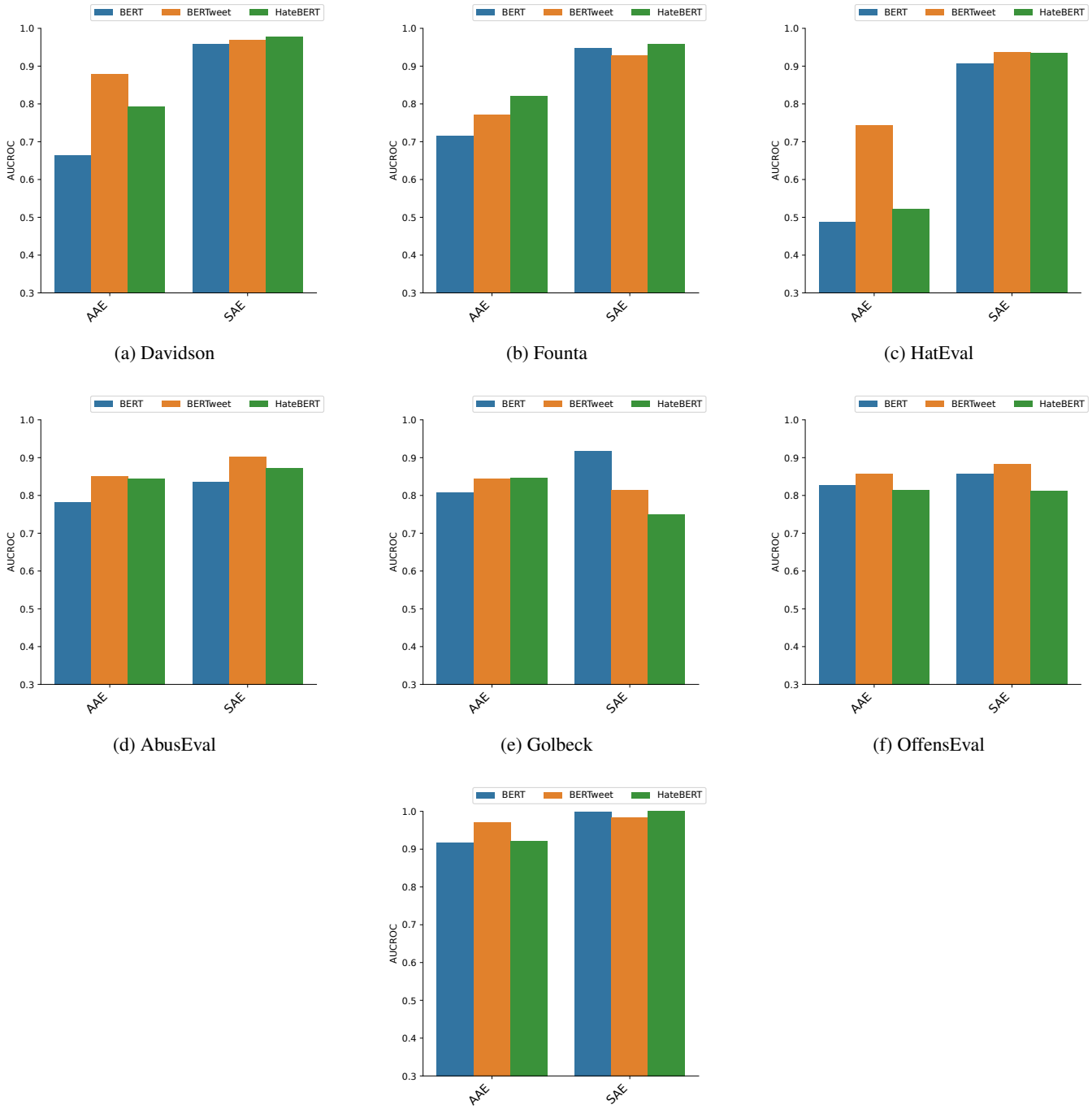


(a) Davidson (b) Founta (c) HatEval

(d) AbusEval (e) Golbeck (f) OffensEval

(g) Waseem

Figure 3: Dialect-wise results for BPSN AUC using GPT-4 general prompt annotated test datasets. Tweets written in AAE are biased toward having more false positives due to the low BPSN AUC scores than tweets written in SAE for all models for Davidson, Founta, HatEval, and AbusEval datasets. There is a change in direction for the Waseem dataset.

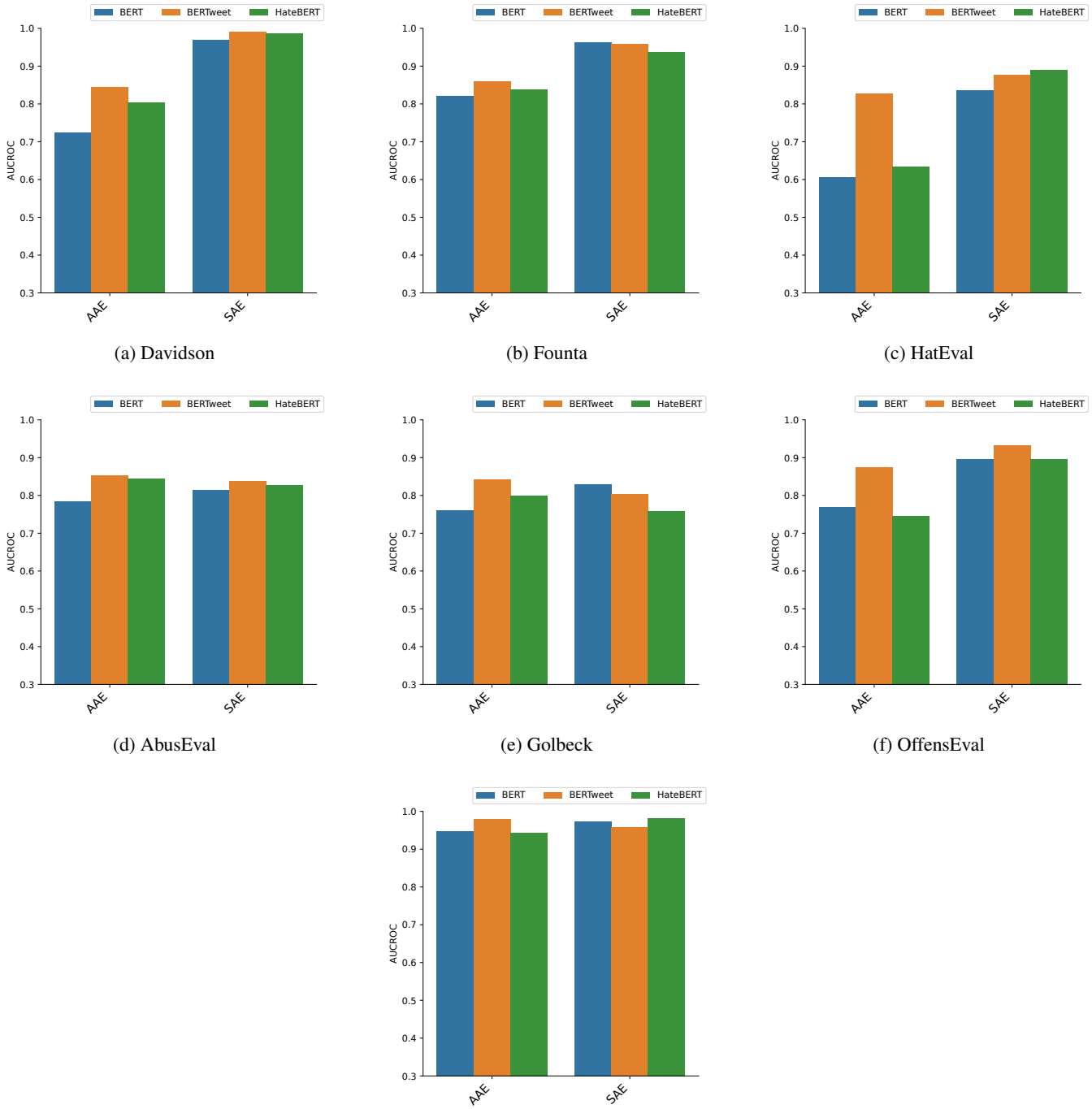


(a) Davidson (b) Founta (c) HatEval

(d) AbusEval (e) Golbeck (f) OffensEval

(g) Waseem

Figure 4: Dialect-wise results for BPSN AUC using GPT-4 few-shot annotation. Results are consistent with the general prompt annotation results except in the Waseem dataset, and the increased difference in BPSN scores of the BERTweet and HateBERT models across the two groups indicates more false positives towards SAE tweets.

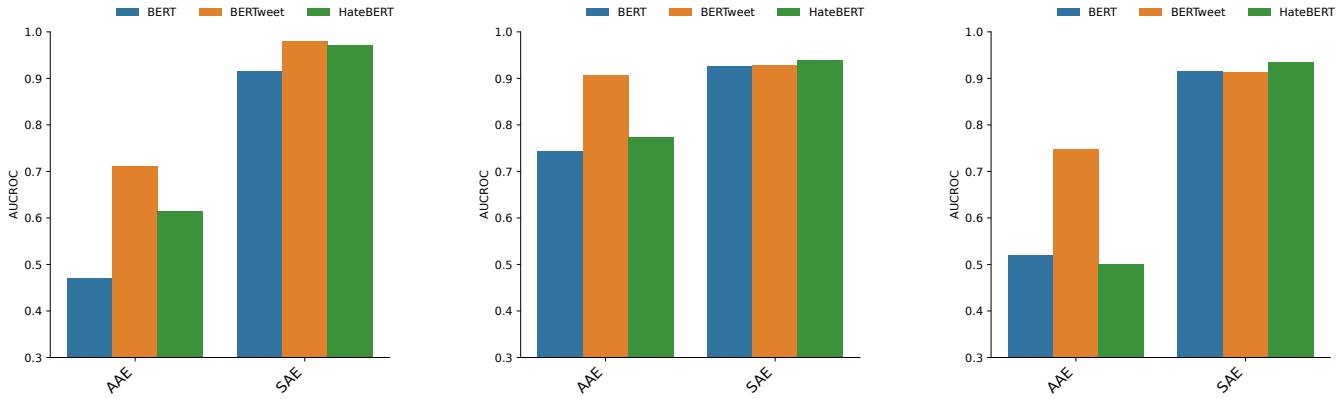


(a) Davidson (b) Founta (c) HatEval

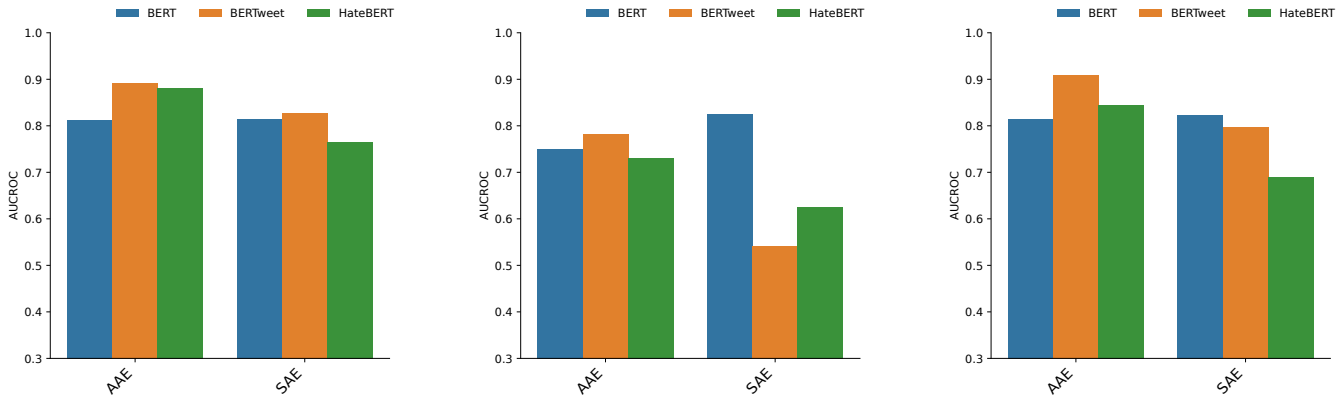
(d) AbusEval (e) Golbeck (f) OffensEval

(g) Waseem

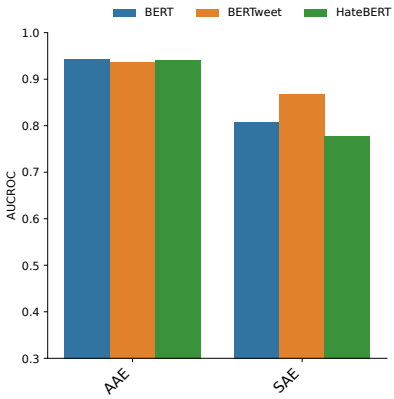
Figure 5: Dialect-wise results for BPSN AUC using GPT-4 chain-of-thought annotation. Results are consistent with general and FS prompt annotation except in the Waseem dataset, where there is no large difference in the BPSN scores of the AAE and SAE tweets, and in the AbusEval dataset.



(a) Davidson (b) Founta (c) HatEval



(d) AbusEval (e) Golbeck (f) OffensEval



(g) Waseem

Figure 6: Dialect-wise results for BPSN AUC using Llama-3 general prompt annotation. Results indicate a bias towards false positives for AAE tweets in all models fine-tuned on the Davison, Founta, and HatEval datasets. SAE tweets are biased towards having false positives in all models fine-tuned on the AbusEval and Waseem datasets.

	GPT-4			Llama-3
Dataset	Gen	FS	CoT	Gen
Davidson	0.385	0.496	0.505	0.253
Founta	0.161	0.209	0.462	0.284
Golbeck	0.288	0.314	0.332	0.184
OffensEval	0.451	0.411	0.434	0.360
HatEval	0.488	0.467	0.569	0.378

Table 35: The Cohen's Kappa scores between GPT-4 annotated datasets using various prompting strategies with dialect priming and human annotation and Llama-3 annotated datasets using general prompt annotation with dialect priming and human annotation.

General prompt	Given the tweet in triple quotes: """Some bad ass females end up with a nigga that is barley attractive cause all you good looking niggas are hoes and worse then bitches""" likely written in African American English. Do you think the tweet is hateful or abusive or normal? Only answer with one of the following: hateful or abusive or normal. Do not provide an explanation for your answer.
Few-shot learning promt	Question: is the tweet in triple quotes """@USER @USER Bitch I wanna know""" likely written in African American English offensive or not offensive? Only respond with offensive or not offensive. Answer: offensive. ... Question: is the tweet in triple quotes """@USER I know! She is still alive....but wondering if her career was railroaded? Odd! Loved her...especially that day!""" likely written in Standard American English offensive or not offensive? Only respond with offensive or not offensive. Answer:
Chain-of-Thought prompt	Question: Is the tweet in triple quotes """@USER @USER @USER @USER @USER @USER I imagine that they weren't honored to be in your presence. If I had to be anywhere near you I would get physically sick""" likely written in Standard American English hateful or abusive or normal? Answer with only one of the following classes: hateful or abusive or normal. The answer should be in the last words of your final sentence after your explanation. Answer: This tweet, likely written in Standard American English, is hateful because it expresses hatred towards the targets and insults the targets @USER, @USER, @USER, @USER, @USER, and @USER. So, the answer is hateful. ... Question: Is the tweet in triple quotes """If I didn't ask for your opinion, then I obviously don't fucking need it""" likely written in Standard American English hateful or abusive or normal? Answer with only one of the following classes: hateful or abusive or normal. The answer should be in the last words of your final sentence after your explanation. Answer:

Table 36: Annotation prompt with dialect priming samples from the Founta (general and CoT) and OffensEval datasets (FS) for the three prompting strategies.