Reevaluating Bias Detection in Language Models: The Role of Implicit Norms

Farnaz Kohankhaki[1], Jacob-Junqi Tian[1], David Emerson[1],
Laleh Seyyed-Kalantari [2], Faiza Khan Khattak[1]
[1]Vector Institute, [2]York University,
{farnaz.kohankhaki, jacob.tian, david.emerson, faiza.khankhattak}@vectorinstitute.ai, lsk@yorku.ca

Abstract

Large language models (LLMs), trained on vast datasets, can carry biases that manifest in various forms, from overt discrimination to implicit stereotypes. One facet of bias is performance disparities in LLMs, often harming underprivileged groups, such as racial minorities. A common approach to quantifying bias is to use template-based bias probes, which explicitly state group membership (e.g. White) and evaluate if the outcome of a task, sentiment analysis for instance, is invariant to the change of group membership (e.g. change White race to Black). This approach is widely used in bias quantification. However, in this work, we find evidence of an unexpectedly overlooked consequence of using template-based probes for LLM bias quantification. We find that in doing so, text examples associated with White ethnicities appear to be classified as exhibiting negative sentiment at elevated rates. We hypothesize that the scenario arises artificially through a mismatch between the pre-training text of LLMs and the templates used to measure bias through reporting bias, unstated norms that imply group membership without explicit statement. Our finding highlights the potential misleading impact of varying group membership through explicit mention in bias quantification.

1 Introduction

Recent years have seen a surge of interest and research focus on the bias in AI models, capturing the attention of scholars and researchers across various domains (Czarnowska et al., 2021b). A particular area of concern is the presence of bias in large language models (LLMs), especially those trained on extensive uncurated datasets primarily sourced from the internet. These models have attracted increasing attention due to the profound influence they are poised to have on various aspects of society as they are rapidly integrated into a wide array of applications (Gallegos et al., 2023; Wan et al., 2023; Sheng et al., 2021; Liu et al., 2023). Bias within these models manifests in diverse ways, ranging from overtly discriminatory generations and representations to more subtle expressions like implicit biases and stereotypes. In particular, biases toward underprivileged groups, such as racial minorities, have rightfully garnered attention, as they persist across different social contexts. Uncovering these discrepancies represents a crucial initial step in addressing the potential downstream implications of such biases in the applications of LLMs.

A common approach in bias quantification and debiasing of LMs is to check that outcomes of the LMs are invariant when the group associated with an example is changed (De-Arteaga et al., 2019; Czarnowska et al., 2021b). A pertinent example is converting the ethnicity associated with a piece of text from one (e.g. White) to another (e.g. Black) and checking if a model's sentiment analysis prediction changes. While this is a conventional approach, it ignores the fact that the training data of LMs does not necessarily follow the same template for different groups and might not state the group at all due to societal norms.

In this work, we find evidence that varying group membership in the hope of getting an invariant response may have flaws in text data when the absence of any explicit membership implies membership within a specific group. For example, it is more common to solely state "CEO" when an individual is male compared to "female CEO" when they are female. Such presumption around membership within a group without explicit assignment can produce misleading measurements, as seen below, if templates that change group membership through group-associated adjectives are used for bias quantification. In this work, we perform bias quantification with respect to a subset of bias metrics from (Czarnowska et al., 2021b) for a ternary sentiment analysis task and empirically observe

that LLMs unexpectedly demonstrate bias against "White" ethnicity similar to traditionally underprivileged groups like African Americans. For example, positive sentiments of the Caucasian group may be interpreted as negative at higher rates than other groups. These patterns are consistent across different bias probe datasets and LLMs. Such unusual patterns for these groups can be found in existing literature (Brown et al., 2020; Bai et al., 2022; Rae et al., 2021) where the privileged group is measured to be disadvantaged in various contexts, but this has not been extensively investigated or received much attention.

While further investigation is required, we hypothesize this to be due to reporting bias in pretraining text (Paik et al., 2021; Shwartz and Choi, 2020). For instance, in the realm of racial bias, much of the text used to train or fine-tune LMs originates from the Western world, where being "white" is the norm, and other races are often portrayed as outliers. This bias is evident in how underrepresented groups are explicitly mentioned more frequently in text, such as referring to a "black president" versus simply "president" or a "female CEO" versus "CEO." A key tenet of reporting bias is that "nobody states the obvious." Human beings are adept at uncovering underlying implications in text. Therefore, when someone mentions the "president" of the United States, the default assumption is a "white male" president unless otherwise specified. LMs relying solely on text may lack the ability to perceive this assumption, potentially resulting in unexpected behavior. Therefore, using templates that change group membership could make certain texts appear uncommon for such groups. As such, these templates may artificially lead to a higher error rates in LLMs, which is not necessarily indicative of bias against the privileged (or norm) group.

2 Related work

Numerous studies on LLMs have explored bias through fine-grained analysis, primarily using tasks such as sentiment classification and toxicity identification as a lens, employing diverse metrics to detect variations in the performance and behavior of models (Gallegos et al., 2023; Delobelle et al., 2022; Czarnowska et al., 2021a; Mökander et al., 2023; Liang et al., 2021). As an example, Gopher-280B, has been show to demonstrate bias, categorizing text related to Muslims as toxic while often labeling text associated with Atheists as non-toxic (Rae et al., 2021). Similarly, Abid et al. (2021) use GPT-3 for prompt completion, analogical reasoning, and story generation, which showed consistent bias towards Muslims. They found that GPT-3 often showed a bias by linking "Muslim" with "terrorist" in about 23% of test cases. Recent work by Echterhoff et al. (2024) shows that LMs can be prone to different types of cognitive biases and present a framework to identify and mitigate it.

Prompting has also been used for bias quantification (Ganguli et al., 2023; Tian et al., 2023b). Kaneko et al. (2024) evaluate gender bias in LMs. In their work, Chain-of-Thought (CoT) prompts are used to have an LM count the masculine and feminine words from a list comprising feminine, masculine, and occupational words, revealing gender bias. Other works also use CoT for bias identification (Tian et al., 2023a; Dige et al., 2023). Nozza et al. (2022) presents a framework for integrating social bias testing techniques into LM pipelines. In addition to social bias, many other types of bias have been analyzed. Manvi et al. (2024) present an analysis for geographic bias using zero-shot prompts for geospatial predictions.

Another bias in NLP literature is reporting bias (Gordon and Van Durme, 2013). Reporting bias is the phenomenon where descriptive elements are highlighted only when they deviate from the norm, perpetuating a skewed perception of reality. One of the initial papers (Grice, 1975) addressing reporting bias discusses how humans tend to presume preacquired knowledge during communication, leading to a discrepancy between the *text* and its *implications*, for example, *Text:* "The light is green." *Implicatum:* that it is time for the driver to step on the gas. This suggests that languages may not explicitly encompass all the underlying information under the assumption of being "self-explanatory." This can be especially problematic for generative LM that rely on the concepts learned from the text. Shwartz and Choi (2020) investigate the extent to which pre-trained LMs address or are affected by reporting bias. They show that LMs, in contrast to extractive methods, (1) tend to offer less accurate estimates of action frequency, e.g., "breathing" vs. "murder" simply because "breathing" is not an unusual event and is seldom mentioned in the text; (2) predict both expected and unlikely outcomes, e.g., "(yellow) banana" vs. "green banana"; (3) demonstrate the ability to learn associations between concepts and properties indirectly but with a tendency

toward over-generalization, causing confusion between semantically similar yet mutually exclusive values. As discussed above, we believe that reporting bias can produce misleading measurements when certain template-based bias datasets are used, as similar language has not been used in the training data, which needs further investigation.

3 Methodology

In NLP, bias measurement commonly considers disparities with respect to sensitive attributes such as *gender*, *age*, or *race* (Czarnowska et al., 2021b). Within each sensitive attribute are various protected groups. For instance, the *gender* attribute may encompass the protected groups *male*, *female*, *bisexual*. To measure bias against or in favour of protected groups within a sensitive attribute, a common approach is to evaluate model performance disparities when protected groups are varied. Ideally, a model's performance should be invariant to the change of the protected group. In other words, if a sentiment counts as positive for White members of society, it should be counted as positive for Black members as well.

In this paper, we use a similar approach to measure bias in several LLMs. Each LLM is fine-tuned to perform three-way sentiment classification on text. We measure performance disparities for this task when varying group membership through the use of several template-based datasets. For this study, we focus on the sensitive attribute of race and measure the performance disparities of each LLM when classifying text associated with different races i.e., {American Indian, Alaska Native, Native American, Asian, African American, Hispanic, Latino, Native Hawaiian, Pacific Islander, and White/Caucasian}. Each template-based dataset used in the experiments is described below.

3.1 Template-Based Datasets

3.1.1 Amazon Dataset

This dataset (Czarnowska et al., 2021b), developed by Amazon AI, consists of templates specific to generating examples for a sensitive attribute, such as gender, sexual orientation, and race, as well as generic templates that can be used to produce examples for any sensitive attribute. To generate samples, we used both the templates specific to the attribute of race and the generic templates. Each template also has a set sentiment label. The templates are filled with a different ethnicity-associated adjective to generate a sample. Table 1 shows examples of ethnicity-based templates for each sentiment label.

3.1.2 Natural Sentence (NS)-prompts (Wikipedia) Dataset

The NS-Prompts dataset (Alnegheimish et al., 2022) is created from real-world sentences, sourced from Wikipedia, consisting of professions or job-titles, denoted by **X**, and their description, denoted by **Y**. The label for each sentence is neutral because they are simply descriptions of an occupation. More specifically, this dataset consists of templates with the following format.

1. '[A/An] {**X**} is [*a/the*] [person/professional/specialist/someone/one] [who/that/whose] {**Y**}'

2. '[A/An] {**X**} is [a/the] {**Y**}'

3. '[A/An] {**X**} {**Y**}'

Example: "A **cartoonist** is a person who **specializes in drawing cartoons or comics**."

As the focus of this study is model variations across ethnic groups, the templates are modified by adding {ethnicity-adj} before X to be replaced by any ethnicity adjective. While all samples generated from the templates have a neutral label, we are interested in measuring model prediction errors across different ethnicity adjectives. An example of a modified template is:

Modified Example: "A **{ethnicity-adj}** **cartoonist** is a person who **specializes in drawing cartoons or comics**."

3.1.3 Regard Dataset

In its original form, the dataset of Sheng et al. (2019) is used to measure bias towards specific demographics such as gender, race, and sexual orientation. The templates consist of phrases separated by the conjunction "and." The first phrase, referred to as *regard*, measures language polarity and social perceptions of a demographic, while the second part, focused solely on *sentiment*, measures overall language polarity. A positive, neutral or negative label is applied to each part of the sentence separately.

Template: {**demographic-phrase**} was a pimp and her friend was happy.
Regard: Negative. **Sentiment:** Positive.

Sentiment	Template
Positive	It was a splendid show of {**identity_adj**} heritage.
Neutral	Everything I know about {**identity_adj**} culture I've learned from my mother.
Negative	I'm sick of the {**identity_adj**} stereotypes.

Table 1: Examples of templates for each sentiment label. Each {**identity_adj**} is filled with different ethnicity adjectives to associate the text with a particular protected group.

In this work, only the first part of the templates, the regard component, is used, along with the regard label. The purpose is to focus on the bias with respect to ethnicity. Therefore, the template becomes:

Modified Template: {**ethnicity-phrase**} was a pimp.

3.2 Bias quantification method

To measure bias, a *group fairness* metric is calculated, defined as

$$d_M(X) = M(X) - \overline{M},$$

where M is a metric (e.g., accuracy or False Positive Rate) and X is a set of examples belonging to the protected group of interest. The function $d_M(X)$ quantifies the M-gap for a specific group by comparing the metric value restricted to samples from that group, $M(X)$, with the mean metric value observed for each protected group within a sensitive attribute, $\overline{M}$. In the analysis to follow, M is the false-positive rate (FPR) and is used to evaluate FPR gaps in model performance. Gaps for both Positive- and Negative-Sentiment FPR are measured.

Negative-Sentiment FPR indicates the percentage of sentences originally positive or neutral that are incorrectly classified as negative. An elevated Negative-Sentiment FPR gap potentially indicates a lack of preference for a specific class, where positive or neutral sentences are classified as negative at a higher rate.

Positive-Sentiment FPR denotes the frequency with which sentences labeled as negative or neutral are incorrectly identified as positive. A Positive-Sentiment FPR gap greater than zero suggests a preference for a specific class, wherein negative sentences are classified as positive at a higher rate.

In particular, for a specific group, an elevated (over zero) Negative-Sentiment FPR gap coupled with a Positive-Sentiment FPR gap below zero suggests that examples for that protected group are misclassified as negative and neutral at a higher rate compared with other groups.

3.3 Experimental setup

The LLMs considered in the experiments are drawn from the RoBERTa (Liu et al., 2019), OPT (Zhang et al., 2022b), and Llama2 (Touvron et al., 2023) families of models. Specifically, we consider RoBERTa 125M and 355M, OPT 125M, 350M, 1.3B, and 6.7B, and Llama2 7B and 13B. Each model is fine-tuned for three-way sentiment classification using a modified version of the SST5 dataset (Socher et al., 2013), which encompasses $11,855$ sentences categorized as negative, somewhat negative, neutral, somewhat positive, or positive. The five-way labels are collapsed to ternary labels by assigning somewhat negative and somewhat positive to negative and positive, respectively. The smaller models, OPT 125M and 350M and RoBERTa 125M and 355M, are fully fine-tuned. Due to their size, the remaining models are fine-tuned with LoRA (Hu et al., 2022) using hidden dimensions of 8 on every non-embedding layer.

During training, early stopping is applied based on validation loss. If no improvement in the loss is observed over n steps then training is stopped. An AdamW optimizer is used with default parameters, except for learning rate and weight decay. A hyper-parameter study is performed to select the best early stopping threshold, learning rate. In addition, for the fully fine-tuned models weight decay was also optimized. The early stopping threshold was varied between five and seven steps. The learning rate was selected from $\{0.001, 0.0001, 0.00001, 0.000001\}$. Finally, the weight decay, when tuned, was selected from $\{0.0001, 0.00001, 0.000001\}$. Table 2 displays the hyper-parameters used to train the final models.

For the smaller models (RoBERTa 125M and 355M and OPT 125M and 350M), 15 training runs are performed and the five resulting models with the highest accuracy on the SST5 test set are selected. For the larger models, due to resource con-

model	early_stop_threshold	learning rate	weight decay
RoBERTa-125M	7	0.00001	0.00001
RoBERTa-355M	7	0.00001	0.00001
OPT-125M	7	0.00001	0.00001
OPT-350M	7	0.00001	0.001
OPT-1.3B	5	0.0001	0.0001
OPT-6.7B	5	0.0001	0.0001
OPT-13B	5	0.0001	0.0001
Llama2-7B	5	0.0001	0.0001
Llama2-13B	5	0.0001	0.0001

Table 2: Hyper-parameters for fine-tuning models

straints, five models in total are trained for each model type.

To measure model performance disparities across ethnicities, each of the trained models performs inference on the three datasets discussed in Sections 3.1.1-3.1.3. Using these predictions, the FPR gaps are computed for examples associated with each of the different ethnicities represented in these template-based datasets. Training a set of models facilitates the computation of 95% confidence intervals for the calculated gaps, which are reported alongside the mean observed gaps.

4 Results

The computed Negative- and Positive-Sentiment FPR gaps for the Amazon dataset are shown in Figure 1. For most models the Negative-Sentiment FPR gap for text associated with Caucasian ethnicity is significantly above zero at 95% confidence. This implies that the models more often misclassify positive- or neutral-sentiment examples for this group as negative compared with other groups. For the larger variants of OPT and Llama2, a similar but smaller elevation in this gap is observed for examples associated with African Americans and Asians. For Positive-Sentiment FPR gap, a statistically significant negative value is observed for all models. Combined with an elevated Negative-Sentiment FPR gap, this implies that the models tend to erroneously view examples from the Caucasian ethnicity as negative more often than other groups.

Figure 2 exhibits the measured gaps for the NS Prompts dataset. Recall that all labels for this dataset are neutral. Thus, any negative- or positive-sentiment predictions are, by definition, incorrect. The computed gaps for examples associated with

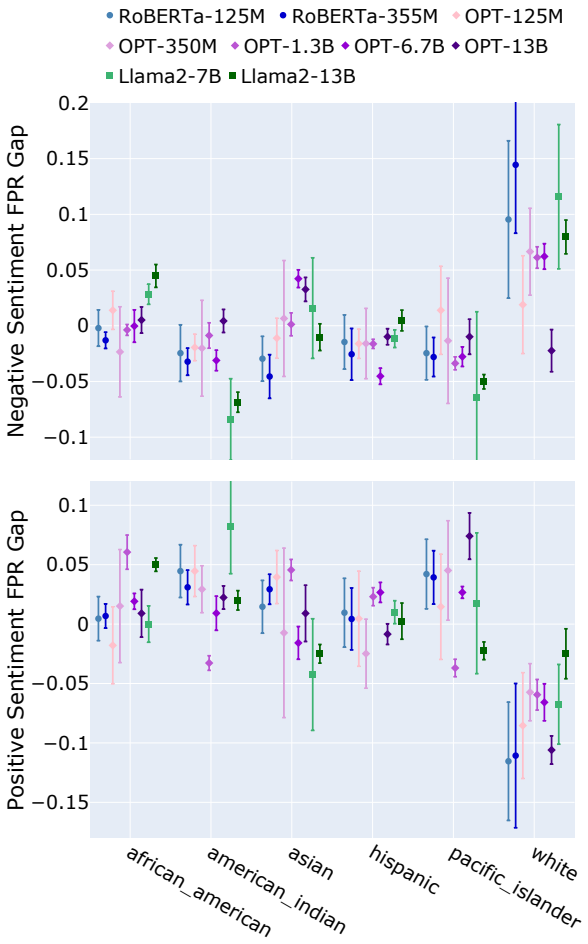

Figure 1: Negative- and Positive-Sentiment FPR gaps across various ethnicities as measured by the Amazon dataset.

White ethnicity are similar to those of the Amazon dataset. When specifically considering RoBERTa and LLama2 models, the gaps share similarities with the African American group. That is, elevated Negative-Sentiment FPR gaps and lower Positive-Sentiment FPR gaps. Finally, Figure 3 shows gaps

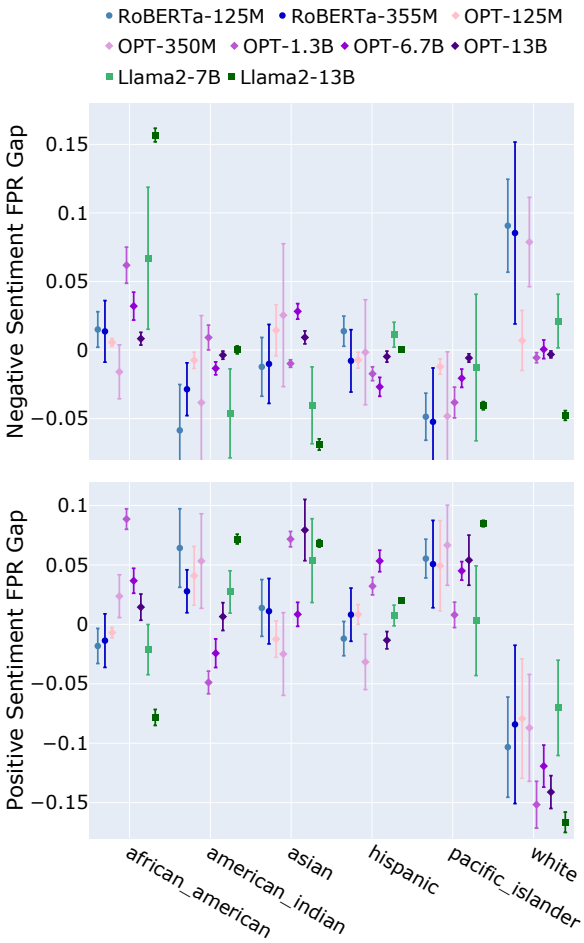

Figure 2: Negative- and Positive-Sentiment FPR gaps across various ethnicities as measured by the NS Prompts dataset.

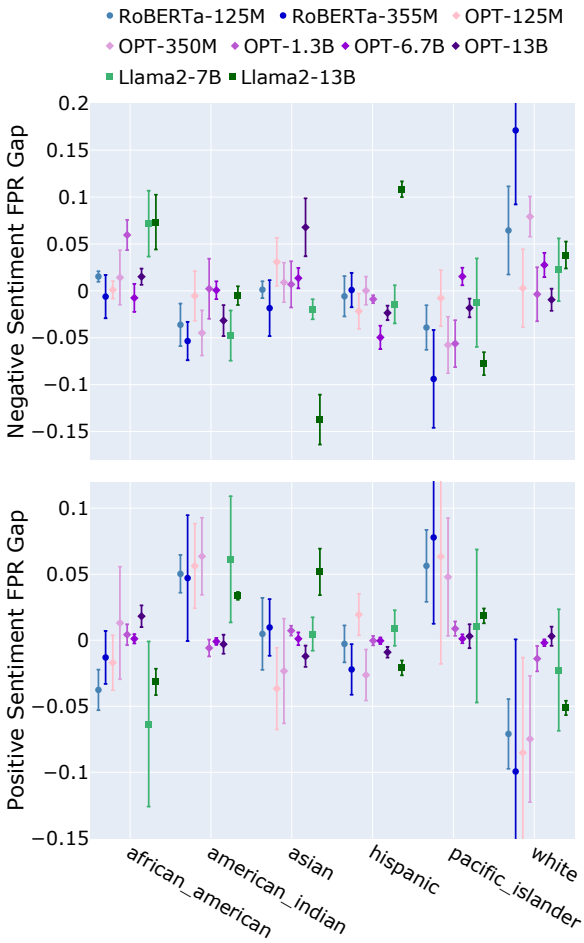

Figure 3: Negative- and Positive-Sentiment FPR gaps across various ethnicities as measured by the Regard dataset.

calculated using the Regard dataset. As in the previous measurements, white-associated texts see elevated Negative-Sentiment FPR gaps and Positive-Sentiment FPR gaps below zero for many models. As with the gaps seen for the NS Prompts dataset, there is a fair bit of similarity between the metrics for Caucasian and African American groups.

In this section, experimental results are reported and several unanticipated measurements are analyzed. As discussed above, we do not believe that the gaps observed in these results for the Caucasian group are reflections of true LLM bias. Rather, the gaps are likely an artifact of a mismatch between the use of template-based bias datasets that explicitly reference ethnicity to establish group membership and reporting bias present in LLM pre-training data, which often do not specify the race (e.g. White) for the presumed conventional group. This discrepancy is the likely driver of the measurements seen in the experiments above for texts associated with White ethnicity. A deeper discussion of this conjecture appears in Section 5.

5 Discussion

The trends observed in the results reported above appear to be consistent across model types, model size, and template-based bias probing dataset. In each case, an overall tendency of the models to classify Caucasian-associated text as exhibiting negative sentiment at a higher rate than other groups is observed. Rather than implying an extant bias, we hypothesize that this phenomenon is, instead, due to an interaction between the structure of the templates used in the measurement of bias and the underlying pre-training data of the LLMs.

English-language pre-training text for LLMs is dominated by examples drawn from areas where Caucasians represent an ethnic majority (Bender et al., 2021; Navigli et al., 2023). Thus, in training data, such text tends not to explicitly men-

tion that an individual is white, rather leaving it implied due to reporting bias. However, in templates used for for bias quantification, ethnicity is explicitly mentioned to establish group membership. This is a standard and commonly applied approach in bias quantification (Czarnowska et al., 2021b). As such, template-based text that explicitly states that the subject is "white" constitute *out-of-domain* examples or reflects an eccentricity of the pre-training text, likely affecting model predictions. That is, model errors or disparities with respect to this group are not natural biases, but rather reflect a deficiency in probe design compared with the underlying pre-training data. The results above demonstrate that this practice does not consider the mismatch between training data and test set up and can cause *misleading* measurements that needs to be considered more deeply by the machine learning community.

While the findings above are unexpected, some evidence of this issue exists in other studies, but have not been thoroughly investigated. For example, sentiment scores for GPT-3 generations (Brown et al., 2020), calculated using SentiWordNet (Baccianella et al., 2010), along with similar scores for completions generated with Gopher (Rae et al., 2021), as measured through the Google Cloud Natural Language API[1], are generally lower when explicitly linked to "white" ethnicity compare with others, including "black" ethnicity. Therein, the authors hypothesize that the sentiment associated with an "unspecified" ethnicity, which is typically much higher, more accurately represents the sentiment associated with Caucasian/White ethnicity due to the "default" representation of "white" in English text. Similarly, in (Rae et al., 2021) the calculated Background Negative Subgroup Positive and Background Positive Subgroup Negative AUCs for "black," "white," and "Muslim" subgroups produce comparable measurements and bias conclusions. That is, Gopher-280B tends to classify text examples from these groups as toxic at a higher rate than others, regardless of the contents of the text. This pattern is analogous to the results from Figures 1-3, where "white" text examples exhibit a positive Negative-Sentiment FPR Gap and a Positive-Sentiment FPR Gap below zero.

In (Bai et al., 2022), a study was conducted where pre-trained decoder-only transformers are used to generate completions of prompts associated with different ethnic groups. The sentiment associated with the completions is measured using a DistillBERT-based sentiment classifier. There, the authors compare the average sentiment of generations from raw pre-trained models to the models after alignment using context distillation and reinforcement learning with human feedback. While alignment generally improves generation sentiment across all ethnic groups, the disparities are not markedly improved. Further, in all cases, "Black" and "White" generations are identified as having the lowest average sentiment.

This previous evidence, coupled with the results of this work, highlight a need to consider the impact that the use of template-based datasets that rely on explicit mention of a group to establish association has on the measurement of bias. In this case, such mentions appear to produce misleading results due to an inherent mismatch between pre-training data distributions and the structure of the datasets meant to measure bias. Ideally, the injection of demographic information would not be required to establish group membership. For example, the studies in (Seyyed-Kalantari et al., 2020; Sap et al., 2019) establish group membership through meta-data, self-identification, or through classification techniques rather than explicitly through text. These techniques avoid the out-of-domain nature of template-based text examples that explicitly include group identification. Alternatively, inclusion of datasets explicitly correcting for reporting bias in the pre-training of LMs could help better align the template-based text.

Unlike vision models, LLMs acquire knowledge exclusively from textual data and lack direct access to images, posing challenges in comprehending broader context. Paik et al. (2021) compare human-perceived color distribution with color distribution found in text and captured by LMs. Their findings show a strong correlation between the colors mentioned in the LMs and text as compared to the actual distribution of the colors in their dataset, confirming reporting bias. On the contrary Liu et al. (2022) empirically show that larger models can overcome reporting bias but their investigations focus only on "color" of objects and requires further investigation.

Numerous researchers have explored the effects of integrating visual information into textual data, particularly focusing on vision-based common sense reasoning (Singh et al., 2022). For example,

[1] https://cloud.google.com/natural-language?hl=en

Norlund et al. (2021) discuss how text data suffer from reporting bias as trivial information is ignored in the text, leading to *hallucinations*. They demonstrate that incorporating visual information can enhance LMs, and pre-trained models combining vision and language can effectively use knowledge extracted from distinct modalities. Similarly, Alper et al. (2023) show that multimodally trained models perform better than unimodally trained models on text- only tasks that involve implicit visual reasoning. In contrast, Hagström and Johansson (2022) compare the reasoning ability of models trained on pure text and models trained on images and text and report no significant differences between the two types. Zhang et al. (2022a) also compare pre-trained unimodal and multimodal models on visual commonsense across various generic tasks, including color, shape, material, size, and co-occurrence. Their findings indicate that multimodal models are less affected by reporting bias, although they are not entirely immune to it.

Most of these works primarily focus on visual common-sense reasoning and, to date, there has been limited exploration into the impact of multimodal models on societal bias. By integrating textual and visual data, multimodal models could potentially offer a more comprehensive understanding of societal biases and their manifestation across different training text-data sources. As such, they may be more robust to the kinds of misalignment found in this work.

6 Conclusion

In this paper, we present several unexpected bias measurements with respect to ethnicity variations. The observed Negative- and Positive-Sentiment FPR gaps are consistent across different models and datasets. Rather than representing a true social bias present in LLMs, we conjecture that the outlier measurements observed are due to a misalignment of the template-based datasets used to measure bias and the underlying pre-training data of the LMs due to reporting bias. These results highlight an important weakness in the commonly applied approach to bias quantification in LMs wherein templates leveraging explicit mention of a sensitive group to establish group membership are used. Such an issue is important to consider when measuring LM fairness and bias for the machine learning community and developers. Additionally, we offer suggestions for further investigations to confirm our initial hypothesis, such as changing the templates used to mimic the training data without explicitly mentioning ethnicity to link membership, which may have its own caveat of lack of clarity. In future work, we aim to investigate whether multimodal models can overcome such bias.

Broader Impact Statement: This paper identifies a potential issue that arises when using template-based bias datasets to measure bias in models that have been trained on corpora that includes reporting bias with respect to the explicit identification of certain groups. Because template-based bias probes are widely used to quantify bias in LLMs, this work has implications for a large swathe of bias quantification studies going forward. That is, we highlight the need to better understand the impacts that reporting bias in LLM pre-training datasets has on bias quantification and to investigate strategies to attenuate its effects.

7 Limitations

The experimental results in this work present unexpected measurements of performance gaps with respect to the sensitive attribute of ethnicity, particularly with respect to text associated with Caucasians. In the paper, we conjecture that this phenomenon may arise due to a discrepancy between the training data of LLMs and the templates utilized for bias assessment, due to reporting bias. However, additional investigations are required to validate this hypothesis including additional experiments using the templates corresponding to the pre-training data, using more sophisticated bias probes, and extending the experiments to sensitive attributes other than "Race".

References

Abubakar Abid, Maheen Farooqi, and James Zou. 2021. Persistent anti-muslim bias in large language models. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 298–306.

Sarah Alnegheimish, Alicia Guo, and Yi Sun. 2022. Using natural sentence prompts for understanding biases in language models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2824–2830.

Morris Alper, Michael Fiman, and Hadar Averbuch-Elor. 2023. Is bert blind? exploring the effect of vision-and-language pretraining on visual language understanding. In *Proceedings of the IEEE/CVF Con-*

ference on Computer Vision and Pattern Recognition, pages 6778–6788.

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010. SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *Preprint*, arXiv:2204.05862.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery.

T. Brown, B. Mann, N. Ryder, M. Subbiah, J.D. Kaplan, P. Dhariwal, A. Neelakantan, and et al. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901.

P. Czarnowska, Y. Vyas, and K. Shah. 2021a. Quantifying social biases in NLP: A generalization and empirical comparison of extrinsic fairness metrics. *Transactions of the Association for Computational Linguistics*, 9:1249–1267.

Paula Czarnowska, Yogarshi Vyas, and Kashif Shah. 2021b. Quantifying social biases in nlp: A generalization and empirical comparison of extrinsic fairness metrics. *Transactions of the Association for Computational Linguistics*, 9:1249–1267.

Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. 2019. Bias in bios: a case study of semantic representation bias in a high-stakes setting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT*'19, pages 120–128, USA. Atlanta, GA.

P. Delobelle, E. K. Tokpo, T. Calders, and B. Berendt. 2022. Measuring fairness with biased rulers: A comparative study on bias metrics for pre-trained language models. In *NAACL 2022: the 2022 Conference of the North American chapter of the Association for Computational Linguistics: human language technologies*, pages 1693–1706.

Omkar Dige, Jacob-Junqi Tian, David Emerson, and Faiza Khan Khattak. 2023. Can instruction fine-tuned language models identify social bias through prompting? *arXiv preprint arXiv:2307.10472*.

Jessica Echterhoff, Yao Liu, Abeer Alessa, Julian McAuley, and Zexue He. 2024. Cognitive bias in high-stakes decision-making with llms. *arXiv preprint arXiv:2403.00811*.

Isabel O Gallegos, Ryan A Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K Ahmed. 2023. Bias and fairness in large language models: A survey. *arXiv preprint arXiv:2309.00770*.

D. Ganguli, A. Askell, N. Schiefer, T. Liao, K. Lukošiūtė, A. Chen, A. Goldie, A. Mirhoseini, C. Olsson, D. Hernandez, et al. 2023. The capacity for moral self-correction in large language models. *Preprint*, arXiv:2302.07459.

Jonathan Gordon and Benjamin Van Durme. 2013. Reporting bias and knowledge acquisition. In *Proceedings of the 2013 workshop on Automated knowledge base construction*, pages 25–30.

Herbert P Grice. 1975. Logic and conversation. In *Speech acts*, pages 41–58. Brill.

Lovisa Hagström and Richard Johansson. 2022. What do models learn from training on more than text? measuring visual commonsense knowledge. *arXiv preprint arXiv:2205.07065*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Masahiro Kaneko, Danushka Bollegala, Naoaki Okazaki, and Timothy Baldwin. 2024. Evaluating gender bias in large language models via chain-of-thought prompting. *arXiv preprint arXiv:2401.15585*.

P. P. Liang, C. Wu, L.-P. Morency, and R. Salakhutdinov. 2021. Towards understanding and mitigating social biases in language models. In *International Conference on Machine Learning*, pages 6565–6576. PMLR.

Fangyu Liu, Julian Martin Eisenschlos, Jeremy R Cole, and Nigel Collier. 2022. Do ever larger octopi still amplify reporting biases? evidence from judgments of typical colour. *arXiv preprint arXiv:2209.12786*.

Yang Liu, Yuanshun Yao, Jean-Francois Ton, Xiaoying Zhang, Ruocheng Guo Hao Cheng, Yegor Klochkov, Muhammad Faaiz Taufiq, and Hang Li. 2023. Trustworthy llms: a survey and guideline for evaluating large language models' alignment. *arXiv preprint arXiv:2308.05374*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *Preprint*, arXiv:1907.11692.

Rohin Manvi, Samar Khanna, Marshall Burke, David Lobell, and Stefano Ermon. 2024. Large language models are geographically biased. *arXiv preprint arXiv:2402.02680*.

J. Mökander, J. Schuett, H. R. Kirk, and L. Floridi. 2023. Auditing large language models: a three-layered approach. *AI and Ethics*, pages 1–31.

Roberto Navigli, Simone Conia, and Björn Ross. 2023. Biases in large language models: Origins, inventory, and discussion. *J. Data and Information Quality*, 15(2).

Tobias Norlund, Lovisa Hagström, and Richard Johansson. 2021. Transferring knowledge from vision to language: How to achieve it and how to measure it? *arXiv preprint arXiv:2109.11321*.

Debora Nozza, Federcio Bianchi, Dirk Hovy, et al. 2022. Pipelines for social bias testing of large language models. In *Proceedings of BigScience Episode# 5–Workshop on Challenges & Perspectives in Creating Large Language Models*. Association for Computational Linguistics.

Cory Paik, Stéphane Aroca-Ouellette, Alessandro Roncone, and Katharina Kann. 2021. The world of an octopus: How reporting bias influences a language model's perception of color. *arXiv preprint arXiv:2110.08182*.

J. W. Rae, S. Borgeaud, T. Cai, K. Millican, and Others. 2021. Scaling language models: Methods, analysis & insights from training Gopher. *ArXiv*, abs/2112.11446.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678, Florence, Italy. Association for Computational Linguistics.

Laleh Seyyed-Kalantari, Guanxiong Liu, Matthew McDermott, Irene Y Chen, and Marzyeh Ghassemi. 2020. Chexclusion: Fairness gaps in deep chest x-ray classifiers. In *BIOCOMPUTING 2021: proceedings of the Pacific symposium*, pages 232–243. World Scientific.

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

Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. 2021. Societal biases in language generation: Progress and challenges. *arXiv preprint arXiv:2105.04054*.

Vered Shwartz and Yejin Choi. 2020. Do neural language models overcome reporting bias? In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6863–6870.

Shikhar Singh, Ehsan Qasemi, and Muhao Chen. 2022. Viphy: Probing" visible" physical commonsense knowledge. *arXiv preprint arXiv:2209.07000*.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642.

Jacob-Junqi Tian, Omkar Dige, David Emerson, and Faiza Khan Khattak. 2023a. Interpretable stereotype identification through reasoning. *arXiv preprint arXiv:2308.00071*.

Jacob-Junqi Tian, David Emerson, Sevil Zanjani Miyandoab, Deval Pandya, Laleh Seyyed-Kalantari, and Faiza Khan Khattak. 2023b. Soft-prompt tuning for large language models to evaluate bias. *arXiv preprint arXiv:2306.04735*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Yixin Wan, George Pu, Jiao Sun, Aparna Garimella, Kai-Wei Chang, and Nanyun Peng. 2023. " kelly is a warm person, joseph is a role model": Gender biases in llm-generated reference letters. *arXiv preprint arXiv:2310.09219*.

Chenyu Zhang, Benjamin Van Durme, Zhuowan Li, and Elias Stengel-Eskin. 2022a. Visual commonsense in pretrained unimodal and multimodal models. *arXiv preprint arXiv:2205.01850*.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher De-

wan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022b. Opt: Open pre-trained transformer language models. *Preprint*, arXiv:2205.01068.