Template-Based Probes Are Imperfect Lenses for Counterfactual Bias Evaluation in LLMs

Anonymous authors
Paper under double-blind review

Abstract

Bias in large language models (LLMs) has many forms, from overt discrimination to implicit stereotypes. Counterfactual bias evaluation is a widely used approach to quantifying bias and often relies on template-based probes that explicitly state group membership. It aims to measure whether the outcome of a task performed by an LLM is invariant to a change in group membership. In this work, we find that template-based probes can introduce systematic distortions in bias measurements. Specifically, we consistently find that such probes suggest that LLMs classify text associated with White race as negative at disproportionately elevated rates. This is observed consistently across a large collection of LLMs, over several diverse template-based probes, and with different downstream task approaches. We hypothesize that this arises artificially due to linguistic asymmetries present in LLM pretraining data, in the form of markedness, (e.g., Black president vs. president) and templates used for bias measurement (e.g., Black president vs. White president). These findings highlight the need for more rigorous methodologies in counterfactual bias evaluation, ensuring that observed disparities reflect genuine biases rather than artifacts of linguistic conventions.

1 Introduction

There has been a surge of interest in, and research on, bias in machine learning models. An important area of focus is the presence of bias in large language models (LLMs), especially those trained on extensive datasets sourced primarily from the internet. These models have attracted increasing attention due to their rapid integration into a wide array of applications (Gallegos et al., 2024; Wan et al., 2023; Sheng et al., 2021; Liu et al., 2023). Bias in these models manifests in diverse ways, ranging from overtly discriminatory generations to more subtle expressions like perpetuating stereotypes. In particular, biases toward underprivileged groups, such as racial minorities, have rightfully garnered attention, as they persist across many social contexts. Uncovering these issues represents a crucial step in addressing the potential implications of such biases in downstream applications.

Counterfactual bias evaluation is a common approach in bias quantification that measures invariance, or lack thereof, in the outcomes of a model for a particular task across different groups, holding all else equal (De-Arteaga et al., 2019; Czarnowska et al., 2021; Martinková et al., 2023; Cimitan et al., 2024). A pertinent example is perturbing the race associated with a piece of text from one group (e.g. White) to another (e.g. Black) and measuring whether a model's sentiment prediction changes. Although this is a widely used approach in bias quantification, it ignores the fact that LLM training data does not necessarily follow the same structure for different groups.

In this work, counterfactual bias quantification experiments are performed spanning several ternary sentiment-analysis tasks. A wide range of LLMs are considered, and two classification techniques, fine-tuning and prompting, are applied to perform the classification tasks. Empirically, we observe clear abnormalities such that LLMs assign disproportionately negative sentiment to texts explicitly associated with White race, similar to traditionally underprivileged groups like African Americans. For example, positive or neutral statements associated with the White group are misinterpreted as negative at higher rates than other groups. These patterns are consistent across bias probing datasets, LLMs, and classification techniques. Overall, the

results demonstrate that template-based bias quantification relying on marking has flaws. These limitations reduce the reliability of such measurements as indicators of actual bias dynamics.

The contributions of this work are summarized as follows.

- We find evidence that counterfactual bias evaluation using template-based probes introduces systematic distortions in bias measurement. The extent to which template-based probes exhibit measurement flaws is systematically quantified through a wide range of experiments. These distortions undermine the usefulness of such datasets as a lens for bias evaluation.
- This paper constructs two new template-based probing datasets from existing work to validate the findings across different domains. These datasets, and the associated techniques for their construction, may be used in future experiments.
- This work provides a strong conjecture as to the underlying cause of the aberrant bias measurements. We hypothesize that these disparities are due to the prevalence of markedness in LLM pretraining text, suggesting new research directions.

2 Related Work

Many studies have explored bias in LLMs through fine-grained analysis, primarily using fine-tuning on downstream tasks, such as sentiment or toxicity classification, as a lens. These studies employ a diverse set of metrics to detect variations in model behavior (Gallegos et al., 2024; Delobelle et al., 2022; Czarnowska et al., 2021; Mökander et al., 2023; Liang et al., 2021; Ribeiro et al., 2020; Levy et al., 2023; Echterhoff et al., 2024; Rae et al., 2021). Standard and Chain-of-Thought (CoT) (Wei et al., 2024) prompting have also been used for bias quantification and identification in LLMs (Ganguli et al., 2023; Cheng et al., 2023; Kaneko et al., 2024; Tian et al., 2023). While some challenges arise in using prompting in this setting (Zayed et al., 2024), it remains a useful tool. Many studies, including those cited above, use template-based probing datasets to perform counterfactual bias analysis in LLMs (Dixon et al., 2018; Huang et al., 2020; Liang et al., 2021; Blodgett et al., 2021; Delobelle et al., 2022; Martinková et al., 2023; Cimitan et al., 2024). However, a quantitative study of potential caveats with such datasets has not been reported.

In Blodgett et al. (2021), a critical study of several bias datasets (StereoSet, CrowS-Pairs, WinoBias, WinoGender) identified systematic issues likely compromising the precision or clarity of biases or stereotyping tendencies of LLMs measured by these datasets. Among other issues, including poor definitions, misalignment, and logical failures, the authors suggest out-of-domain text due to markedness as potentially clouding the proposed measurements. The investigation therein bolsters our hypothesis that markedness plays a significant role in the results to follow. However, their study does not quantify the effect of these flaws. Rather, it simply identifies qualities that may be problematic. Their work also focuses on entirely different datasets than those studied here. Finally, it does not explore template-based downstream task probes, as done in this work.

Several studies considering the extent to which markedness or reporting bias are incorporated into LLMs or affects their predictions exist (Bender et al., 2021; Wolfe & Caliskan, 2022b;a; Cheng et al., 2023; Shwartz & Choi, 2020). Each of these studies notes that markedness plays a critical role in the way models make predictions and that these models have internalized aspects of markedness through their training. These studies reveal certain biases related to markedness or reporting bias but do not investigate counterfactual bias or template-based probes.

3 Methodology

In natural language processing, bias measurement typically examines disparities in sensitive attributes such as *gender* or *race* (Czarnowska et al., 2021). Each attribute includes various protected groups. Herein, the attribute of race is specifically considered. Within the sensitive attribute of *race*, we restrict our focus to the protected groups of *American Indian*, *Asian*, *African American*, *Hispanic*, *Pacific Islander*, and *White*.

A standard bias measurement approach evaluates model performance disparities when protected groups are varied. Ideally, model performance remains invariant to group changes or substitutions.

It should be noted that race and ethnicity have distinct anthropological definitions, yet many studies and bias datasets use the terms interchangeably, including those used in the experiments to follow. For instance, the templates in Czarnowska et al. (2021), discussed below, categorize "Hispanic" under race, though it is commonly considered an ethnicity (Lopez et al., 2023). To maintain consistency with prior work, the term "race" is used throughout, despite its imperfect fit for some protected groups.

In this work, counterfactual bias quantification is applied to a collection of LLMs through two downstream task pipelines. In the first, LLMs are fine-tuned for three-way sentiment classification using the SST5 dataset (Socher et al., 2013), and bias is measured by varying group membership across multiple template-based datasets. In the second, LLMs classify template-based datasets directly, without fine-tuning, through prompting. As this study examines race as the sensitive attribute, we measure classification performance disparities across racial groups. Both pipelines analyze false positive rate (FPR) discrepancies between groups. Three template-based datasets are used and detailed in the sections to follow.[1]

3.1 Template-Based Datasets

3.1.1 Amazon Dataset

This dataset consists of templates for generating examples for a specific sensitive attribute, such as gender and race, as well as generic templates that may be used to produce examples for any sensitive attribute (Czarnowska et al., 2021). Templates specific to the attribute of race and generic templates are both used for the experiments. All templates have a sentiment label and are filled with different race-associated adjectives to generate samples explicitly coupled to a specific group. Examples are as follows.

(Positive) It was a splendid show of {**race_adj**} heritage.

(Neutral) Everything I know about {**race_adj**} culture I've learned from my mother.

(Negative) I'm sick of the {**race_adj**} stereotypes.

3.1.2 NS-Prompts Dataset

The NS-Prompts dataset is created from templates using sentences sourced from Wikipedia consisting of professions or job titles and their description (Alnegheimish et al., 2022). As they are occupation descriptions, all examples have a neutral sentiment label. The original templates have the following potential formats.

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

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

[A/An] {X} {Y}.

where X is a profession or job title and Y is the associated description. A completed example might be "A cartoonist is a person who specializes in drawing cartoons or comics."

The focus here is on model variations across racial groups. Thus, the templates are modified by adding {race_adj} before X to be filled with a race adjective tying examples to a specific group. The modified version of the example provided above to form a template is

"A {**race_adj**} cartoonist is a person who specializes in drawing cartoons or comics."

Though all samples have a neutral label, we are interested in measuring model prediction errors while varying these group adjectives.

[1] All experimental code is found at: *Github link withheld for double-blind submission.*

3.1.3 Regard Dataset

The dataset of Sheng et al. (2019) measures bias toward specific demographics, including race. Its templates consist of phrases joined by "and." The first phrase, *regard*, assesses language polarity and social perceptions of a demographic, while the second, focused on *sentiment*, measures overall language polarity. Each phrase receives a positive, neutral, or negative label. An original template highlighting regard and sentiment phrases in red and blue, respectively, is shown below.

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

Labels: **Regard:** Negative. **Sentiment:** Positive.

In this work, only the first part of the templates, corresponding to the regard component, is used, along with the regard label. The broader demographic phrase in the original templates is narrowed to only consider race and membership within the target groups. As an example, the template above becomes:

Modified Template: {**race_phrase**} was a pimp.

3.2 Bias Quantification Metric

The bias metric in the experiments is defined as

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

where M is a performance metric 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, $\overline{M}$. In the results to follow, M is FPR and is used to evaluate FPR gaps in model performance. Gaps for both Positive- and Negative-Sentiment FPR are measured. Mean gaps and 95% confidence intervals (CIs) are calculated based on five runs.

Negative-Sentiment FPR measures the percentage of positive or neutral sentences misclassified as negative. An elevated Negative-Sentiment FPR gap suggests a potential lack of preference for a group, where such sentences are classified as negative more often. Conversely, Positive-Sentiment FPR denotes the rate at which negative or neutral sentences are misclassified as positive. A Positive-Sentiment FPR gap above zero suggests a preference for a group, where negative or neutral sentences are classified as positive more frequently. An elevated Negative-Sentiment FPR gap combined with a Positive-Sentiment FPR gap below zero indicates that a group's examples are classified as negative or neutral more often than others, suggesting the group is viewed unfavorably by the LLM.

3.3 Fine-Tuning Experimental Setup

The LLMs considered in this set of experiments are drawn from the RoBERTa (Liu et al., 2020), OPT (Zhang et al., 2022), Llama-2/3 (Touvron et al., 2023), and Mistral (Jiang et al., 2023) families of models. Specifically, RoBERTa 125M and 355M, OPT 125M, 350M, 1.3B, and 6.7B, Llama-2 7B and 13B, Llama-3 8B, and Mistral 7B are considered. Each model is fine-tuned for three-way sentiment classification using a modified version of the SST5 dataset, 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. 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). Each model is trained five separate times with different random seeds. Detailed hyperparameter settings for fine-tuning are included in Appendix A.

To measure model performance disparities across races, each of the trained models performs inference on examples generated from the three datasets discussed in Sections 3.1.1-3.1.3 to predict their sentiment. Using these predictions, FPR gaps are computed for examples associated with the different racial groups. Training a set of models facilitates the computation of 95% CIs for the gaps, which are reported alongside the mean gaps.

3.4 Prompting Experimental Setup

Three prompting strategies are applied to predict sentiment. These are zero-shot prompts, 9-shot prompts with shots drawn from two sentiment analysis datasets, and zero-shot CoT prompts (Kojima et al., 2024). For all prompting experiments, Hugging Face's text-generation pipeline is used for the base models of OPT-6.7B, Llama-2-7B, Llama-3-8B, and Mistral-7B. These models correspond to the Hugging Face identifiers `facebook/opt-6.7b`, `meta-llama/Llama-2-7b-hf`, `meta-llama/Meta-Llama-3-8B`, and `mistralai/Mistral-7B-v0.1`. Sampling is turned on, and a temperature of 0.8 is used for all generations, including reasoning traces. Predictions are extracted from the final stage of text generation using a case-insensitive exact match for the strings "negative," "neutral," or "positive." The first instances of such a match are taken as the predicted label. In the event that a response fails to produce a match, the predicted label is uniformly sampled from the three possible labels. In all but the reasoning generation stage of zero-shot CoT, models produce a maximum of three tokens in their response.

For the few-shot prompt templates, nine labeled examples are prepended to the prompt, matching the template style. Two distinct experiments are conducted with labeled demonstrations drawn from either the SST5 or SemEval (Mohammad et al., 2018) datasets. For SST5, labels are collapsed in the same way described in Section 3.3. The SemEval polarities are condensed via the mapping {*Negative*: [-3, -2], *Neutral*: [-1, 0, 1], *Positive*: [2, 3]}. In both cases, to avoid any few-shot bias (Gupta et al., 2024), demonstrations are balanced between negative, neutral, and positive (3 each), but order is random. Demonstrations are constant across models, but are resampled across the five prediction runs of each experiment. For reproducibility, random seeds for demonstration selection and all generations, including other prompts, are set to {2024, 2025, 2026, 2027, and 2028} across the five runs.

The final prompting approach, zero-shot CoT, uses two sequential prompt templates. CoT prompting is not applied to OPT, as the model has limited reasoning capacity (Liang et al., 2023). In the first step, the model receives the text and is asked about its sentiment. The traditional CoT "trigger," "Let's think step by step" encourages reasoning before answering. Reasoning traces are capped at 64 tokens. To quantify generation stochasticity, each example is predicted five times. All prompt templates for each of the prompting strategies and other settings appear in Appendix B.

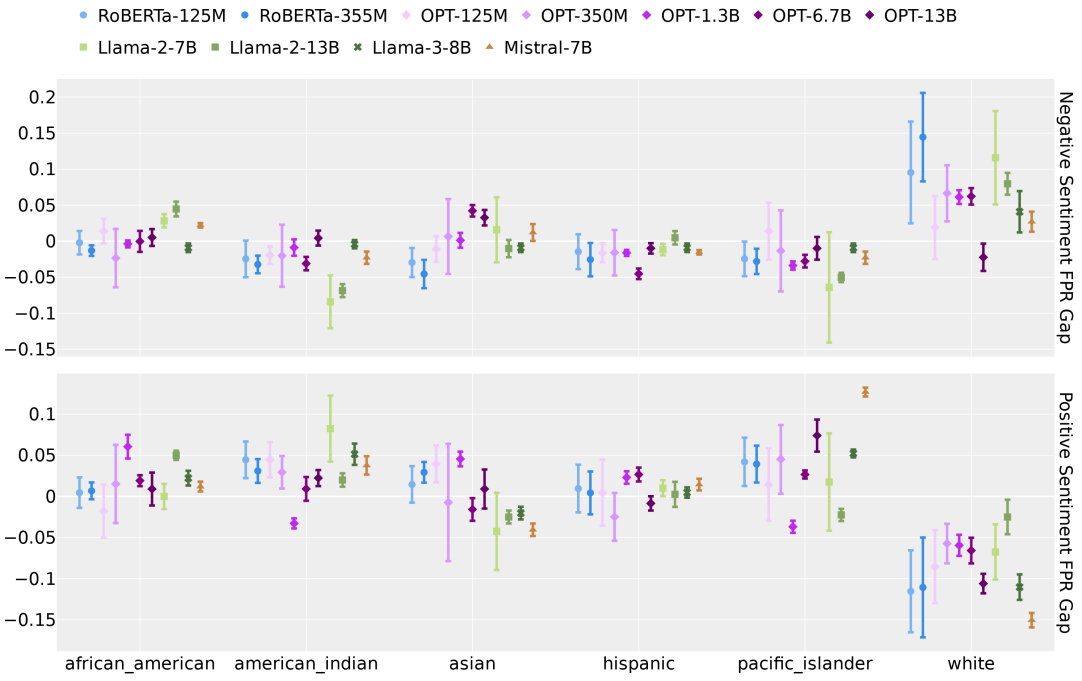

Figure 1: Negative- and Positive-Sentiment FPR gaps as measured by the Amazon dataset.

4 Results

4.1 Fine-Tuning Results

The Negative- and Positive-Sentiment FPR gaps for the Amazon dataset are shown in Figure 1. For most models, the Negative-Sentiment FPR gap for White-associated text 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 others. For large OPT, Llama-2 and Mistral LLMs, a similar but smaller elevation in this gap is observed for examples associated with African Americans and Asians. For the Positive-Sentiment FPR gap, a significant negative value is observed for all models. More recent models, Llama-3 and Mistral, exhibit some of the largest negative gaps. Combined with an elevated Negative-Sentiment FPR gap, this implies that the models tend to view examples from the White group in a negative light more often than other groups.

Figure 2 displays the measured gaps for the NS-Prompts dataset. Recall that all labels for this dataset are neutral. Thus, any non-neutral predictions are, by construction, incorrect. When considering RoBERTa and Llama-2 models, the identified gaps share similarities with the African-American group. That is, elevated Negative-Sentiment FPR gaps and Positive-Sentiment FPR gaps below zero. While the negative-sentiment FPR gaps for other models are near zero for White examples, all models produce negative and statistically significant Positive-Sentiment FPR gaps. This implies that neutral examples associated with White race are construed as positive at much lower rates relative to other groups.

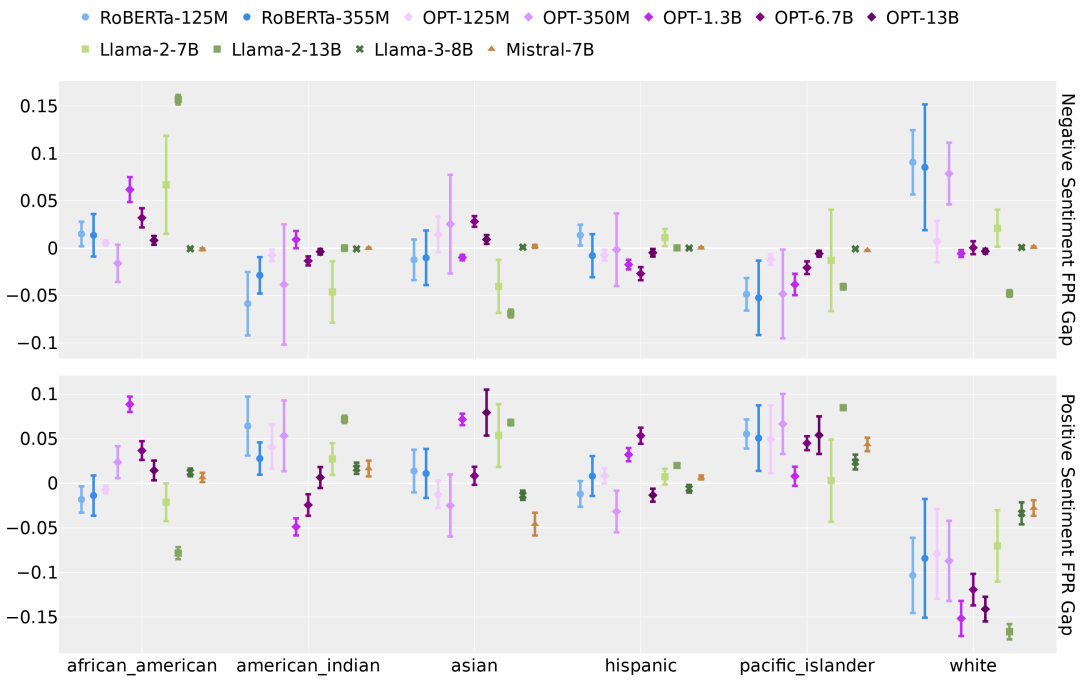

Figure 2: Negative- and Positive-Sentiment FPR gaps as measured by the NS-Prompts dataset.

Results for the Regard dataset reveal similar trends to the Amazon and NS-Prompts experiments. However, the gaps, displayed in Figure 3, are somewhat smaller. As in previous measurements, White-associated texts see elevated Negative-Sentiment FPR gaps and Positive-Sentiment FPR gaps below zero for many models. Furthermore, strong parallels exist for the gaps observed for text associated with African Americans. This is especially true for RoBERTa, small OPT, Llama-2, and Llama-3 models, where the gaps for these groups are highly correlated.

The measurements in these results are surprising. However, as discussed in detail in Section 5 below, the gaps observed for the White group are not believed to be reflections of true bias. Rather, we conjecture that

they are an artifact of a mismatch between the template-based probing datasets that explicitly reference race to link membership and markedness in LLM pretraining data.

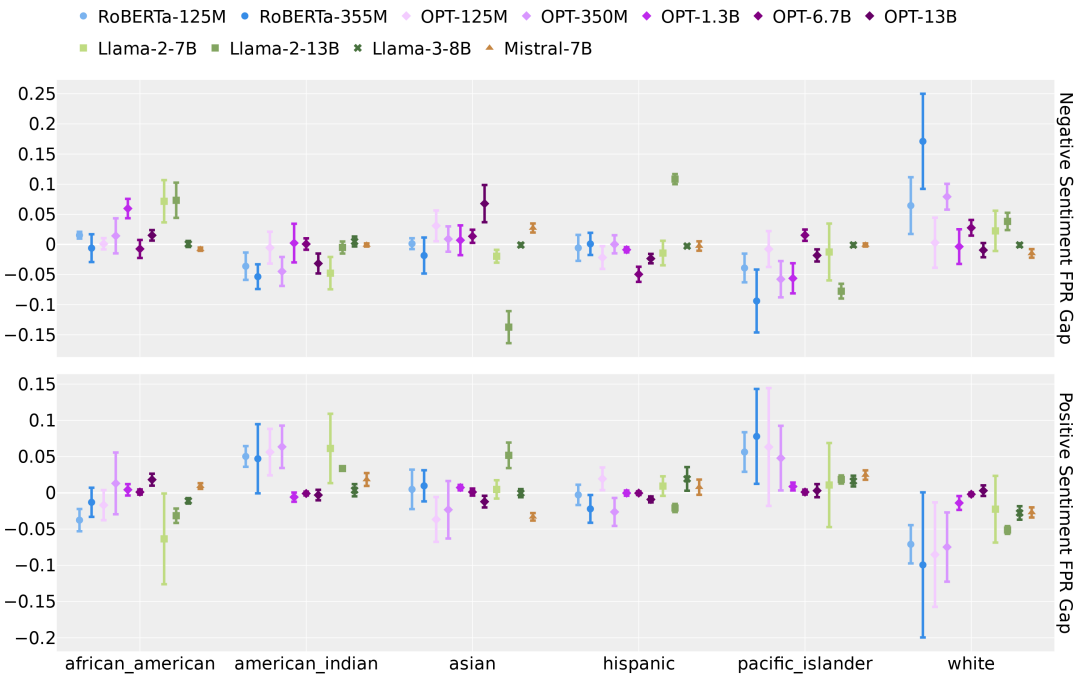

Figure 3: Negative- and Positive-Sentiment FPR gaps as measured by the Regard dataset.

4.2 Prompt-Based Results

The results in Section 4.1 exhibit clear anomalies when measuring performance gaps using template-based probes. A natural question is whether such irregularities arise due to the task-specific fine-tuning step or represent an intrinsic quality of the LLMs. To further isolate the issue to LLM pretraining, prompting is used to perform sentiment classification for the Amazon dataset, shedding the need for fine-tuning. The experiments are limited to decoder-only models of sufficient size to ensure that classification performance adequately exceeds that of a random classifier.

The average classification accuracy of the prompting and fine-tuning approaches on the Amazon dataset is reported in Appendix C. Generally, the accuracy of prompt-based classification is lower than the fine-tuning counterpart. This is especially true for the oldest model, OPT. The best performing method is the 9-shot prompt drawn from SST5 with an accuracy of 71.6% using Llama-3-8B. Many fine-tuned models approach or outperform this accuracy. Nonetheless, as classifiers, the prompted LLMs perform well above a random model. Perhaps due to model size, reasoning in the form of zero-shot CoT does not significantly improve performance (Wei et al., 2024).

As in Section 4.1, Negative- and Positive-Sentiment FPR gaps are computed for each LLM's predictions. These gaps are exhibited in Figure 4. Due to the lower accuracy and generation volatility, the gap CIs are visibly wider than those of the fine-tuning experiments. Nonetheless, a clear and familiar pattern is seen in these results. Positive mean gaps in Negative-Sentiment FPR are present across nearly all examples for African American and White races. Similarly, negative mean gaps for Positive-Sentiment FPR are measured for both races in most settings. The consistency between these results and those of the fine-tuning experiments strongly suggests that the irregularities present in the template-based measurements are not the result of the fine-tuning stage, but are, rather, an expression of an intrinsic aspect of the LLMs themselves.

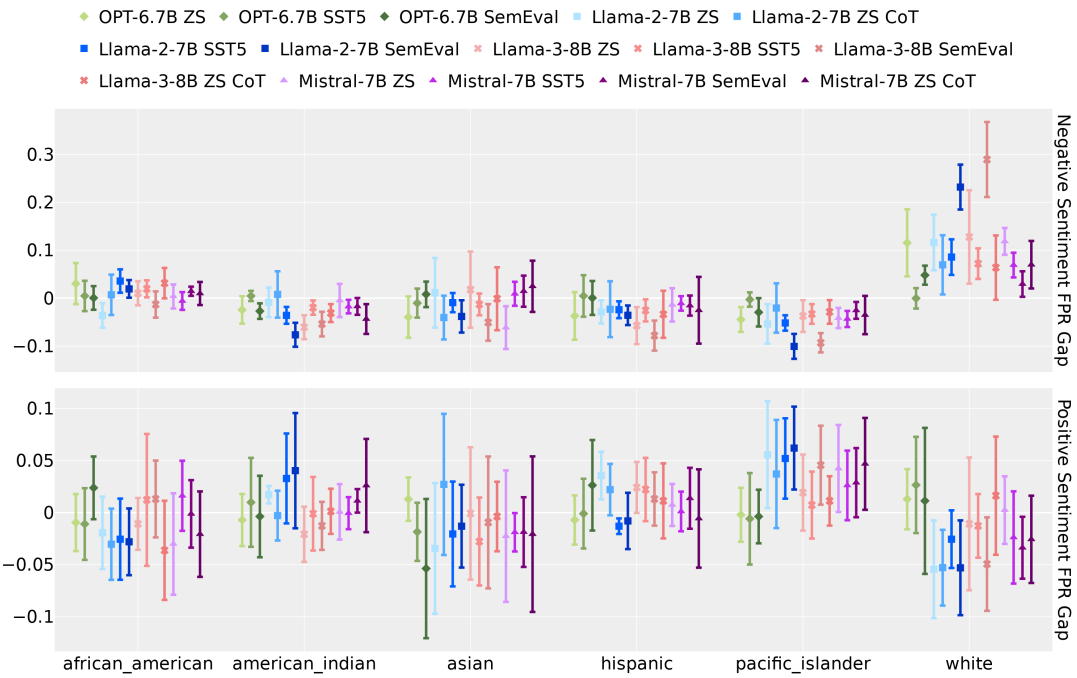

Figure 4: Negative- and Positive-Sentiment FPR gaps as measured by the Amazon dataset with prompt-based classification. In the legend, ZS stands for zero-shot. SST5 and SemEval indicate 9-shot prompts with examples drawn from those datasets.

5 Discussion

Across the experiments an overall tendency of the models to classify White-associated text as exhibiting negative sentiment at a higher rate than other groups is observed. The trends in the results above are consistent between model type, model size, template-based probing dataset, and even classification strategy. The overall agreement of the prompting and fine-tuning results indicates that the observed gaps are not linked to idiosyncrasies in the fine-tuning process but are, rather, more fundamental to the LLMs themselves and the design of the template-based probes. In addition, the models chosen for experimentation are base versions. That is, their predictions are not influenced by interceding alignment techniques (Bai et al., 2022; Rafailov et al., 2023), which might otherwise obscure behavior learned during pretraining. Rather than implying an extant bias, we hypothesize below that this phenomenon is due to an interaction between the structure of the templates used in the measurement of bias and markedness in LLM pretraining data. Regardless of the underlying cause, these observations should lead us to re-think the clarity of counterfactual bias analysis in this context.

5.1 Markedness and Template-Based Probes

The concept of default group membership in the absence of direct assignment has been extensively studied in linguistics under the category of markedness (Trubetzkoy, 1969; Jakobson, 1972; Comrie, 1986). In sociological contexts, markedness considers the linguistic differences that arise when referring to default groups compared to others. The concept was first extended to social categories, such as gender and race, in Waugh (1982) wherein it is noted that U.S. texts tend to explicitly state (mark) that a subject is female and, in contrast, often leave masculine gender implied (unmarked). That is, it is more common to use the term "CEO" when an individual is male compared to "female CEO" when they are female. Many subsequent studies have affirmed that markedness extends to race and, in particular, that non-White individuals are often referred to along with their race, while White race membership tends to go unstated (Cheryan & Markus, 2020; Berkel et al., 2017; Brekhus, 2002).

English pretraining data for LLMs is dominated by text drawn from areas where the racial majority is White (Bender et al., 2021; Navigli et al., 2023). Several studies have confirmed that markedness is widespread in internet data, with White race and male gender constituting the unmarked defaults (Wolfe & Caliskan, 2022a; Bailey et al., 2022). Furthermore, it has been shown that models, and LLMs in particular, trained on web data reflect these markedness characteristics (Bender et al., 2021; Wolfe & Caliskan, 2022b;a). On the other hand, in templates commonly used for bias quantification, race is explicitly mentioned to establish group membership. As such, template-based text that explicitly establishes that the subject is "White" essentially constitute out-of-domain examples (Blodgett et al., 2021; Dressler, 1985). Such a mismatch likely influences model predictions.

Although further investigation is required, we hypothesize that the disparities observed in Section 4 associated with the White group are due to the prevalence of markedness in LLM pretraining text. A key assumption underlying unmarked representations is that humans are adept at recognizing unstated implications in text. LLMs trained solely on unstructured next-token prediction, which underpins almost all modern LLM pretraining, may lack the ability to perceive such implications, resulting in surprising behavior. Using templates that represent group membership through explicit description likely makes certain text appear uncommon for traditionally unmarked groups. As such, these templates may lead to artificially elevated error rates in LLMs, skewing bias measurements in unpredictable ways and clouding the lens provided by datasets of this structure.

Including datasets that explicitly correct for markedness in LLM pretraining could better align template-based text. Appendix D suggests that more recent LLMs, trained on larger multilingual datasets, show improvements in measured gap sizes. Both Llama-3-8B and Mistral-7B have the smallest difference between the most positive and negative gaps for Negative-Sentiment FPR, averaged over the three datasets. Llama-3-8B also produces the lowest average difference for Positive-Sentiment FPR. Given that White-group gaps often rank among the extremes, this suggests newer models may be less affected by markedness.

6 Conclusions and Future Work

This paper presents unexpected, and likely flawed, bias measurements related to race when using template-based bias probes. The measurements remain consistent across a number of different experimental settings and varied datasets. Rather than indicating genuine social bias in the LLMs, we conjecture that these outliers stem from a misalignment between template-based bias probes and LLM pretraining data due to markedness. Regardless of the underlying cause, these findings highlight the need to consider the impact that the use of bias probes relying on marked text has on the measurement of bias. In this case, such probes produce largely misleading results. Ideally, artificial injection of demographic information would not be required. For example, the studies of Seyyed-Kalantari et al. (2020) and Sap et al. (2019) establish group membership through meta-data, self-identification, or classification techniques rather than explicitly in text. These methods avoid the out-of-domain nature of template-based examples of the kind studied here and do not see the unnatural patterns we observed. Future work will design experiments to validate the misalignment due to markedness conjecture and construct straightforward ways to mitigate such issues in LLMs.

References

Sarah Alnegheimish, Alicia Guo, and Yi Sun. Using natural sentence prompts for understanding biases in language models. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz (eds.), *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 2824–2830. Association for Computational Linguistics, Seattle, United States, July 2022. doi: 10.18653/v1/2022.naacl-main.203. URL https://aclanthology.org/2022.naacl-main.203.

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. Training a helpful and harmless assistant with reinforcement learning from human feedback, 2022. Preprint at https://arxiv.org/abs/2204.05862.

April H. Bailey, Adina Williams, and Andrei Cimpian. Based on billions of words on the internet, people=men. *Science Advances*, 8(13):eabm2463, 2022. doi: 10.1126/sciadv.abm2463. URL https://www.science.org/doi/abs/10.1126/sciadv.abm2463.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 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, pp. 610–623. Association for Computing Machinery, New York, NY, USA, 2021. ISBN 9781450383097. doi: 10.1145/3442188.3445922. URL https://doi.org/10.1145/3442188.3445922.

Laura Van Berkel, Ludwin E. Molina, and Sahana Mukherjee. Gender asymmetry in the construction of american national identity. *Psychology of Women Quarterly*, 41:352–367, 2017.

Su Lin Blodgett, Gilsinia Lopez, Alexandra Olteanu, Robert Sim, and Hanna Wallach. Stereotyping Norwegian salmon: An inventory of pitfalls in fairness benchmark datasets. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 1004–1015. Association for Computational Linguistics, Online, August 2021. doi: 10.18653/v1/2021.acl-long.81. URL https://aclanthology.org/2021.acl-long.81.

Wayne Brekhus. A sociology of the unmarked: Redirecting our focus. *Sociological Theory*, 16(1):34–51, 2002.

Myra Cheng, Esin Durmus, and Dan Jurafsky. Marked personas: Using natural language prompts to measure stereotypes in language models. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1504–1532. Association for Computational Linguistics, Toronto, Canada, July 2023. doi: 10.18653/v1/2023.acl-long.84. URL https://aclanthology.org/2023.acl-long.84.

Sapna Cheryan and Hazel Rose Markus. Masculine defaults: Identifying and mitigating hidden cultural biases. *Psychological Review*, 127(6):1022—-1052, 2020.

Ana Cimitan, Ana Alves Pinto, and Michaela Geierhos. Curation of benchmark templates for measuring gender bias in named entity recognition models. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue (eds.), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 4238–4246, Torino, Italia, May 2024. ELRA and ICCL. URL https://aclanthology.org/2024.lrec-main.378/.

Bernard Comrie. Markedness, grammar, people, and the world. In Fred R. Eckman, Edith A. Moravcsik, and Jessica R. Wirth (eds.), *Markedness*, pp. 85–106. Springer US, Boston, MA, 1986. ISBN 978-1-4757-5718-7. doi: 10.1007/978-1-4757-5718-7_6. URL https://doi.org/10.1007/978-1-4757-5718-7_6.

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

Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. 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, pp. 120–128. Association for Computing Machinery, USA, 2019. Atlanta, GA.

Pieter Delobelle, Ewoenam Tokpo, Toon Calders, and Bettina Berendt. Measuring fairness with biased rulers: A comparative study on bias metrics for pre-trained language models. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz (eds.), *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1693–1706. Association for Computational Linguistics, Seattle, United States, July 2022. doi: 10.18653/v1/2022.naacl-main.122. URL https://aclanthology.org/2022.naacl-main.122.

Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '18, pp. 67–73, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450360128. doi: 10.1145/3278721.3278729. URL https://doi.org/10.1145/3278721.3278729.

Wolfgang U. Dressler. On the predictiveness of natural morphology. *Journal of Linguistics*, 21(2):321–337, 1985. ISSN 00222267, 14697742. URL http://www.jstor.org/stable/4175791.

Jessica Echterhoff, Yao Liu, Abeer Alessa, Julian McAuley, and Zexue He. Cognitive bias in high-stakes decision-making with LLMs, 2024. Preprint at https://arxiv.org/abs/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. Bias and fairness in large language models: A survey. *Computational Linguistics*, 50(3):1097–1179, 09 2024. ISSN 0891-2017. doi: 10.1162/coli_a_00524. URL https://doi.org/10.1162/coli_a_00524.

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

Karan Gupta, Sumegh Roychowdhury, Siva Rajesh Kasa, Santhosh Kasa, Anish Bhanushali, Nikhil Pattisapu, Prasanna Srinivasa Murthy, and Alok Chandra. How robust are llms to in-context majority label bias? In *AAAI 2024 Workshop on Responsible Language Models*, 2024.

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

Po-Sen Huang, Huan Zhang, Ray Jiang, Robert Stanforth, Johannes Welbl, Jack Rae, Vishal Maini, Dani Yogatama, and Pushmeet Kohli. Reducing sentiment bias in language models via counterfactual evaluation. In Trevor Cohn, Yulan He, and Yang Liu (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 65–83, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.7. URL https://aclanthology.org/2020.findings-emnlp.7/.

Roman Jakobson. Verbal communication. *Scientific American*, 227(3):72–81, 1972. ISSN 00368733, 19467087. URL http://www.jstor.org/stable/24927429.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, 2023. Preprint at https://arxiv.org/abs/2310.06825.

Masahiro Kaneko, Danushka Bollegala, Naoaki Okazaki, and Timothy Baldwin. Evaluating gender bias in large language models via chain-of-thought prompting, 2024. Preprint at https://arxiv.org/abs/2401.15585.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22. Curran Associates Inc., Red Hook, NY, USA, 2024. ISBN 9781713871088.

Sharon Levy, Neha John, Ling Liu, Yogarshi Vyas, Jie Ma, Yoshinari Fujinuma, Miguel Ballesteros, Vittorio Castelli, and Dan Roth. Comparing biases and the impact of multilingual training across multiple languages. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 10260–10280. Association for Computational Linguistics, Singapore, December 2023. doi: 10.18653/v1/2023.emnlp-main.634. URL https://aclanthology.org/2023.emnlp-main.634.

Paul Pu Liang, Chiyu Wu, Louis-Philippe Morency, and Ruslan Salakhutdinov. Towards understanding and mitigating social biases in language models. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 6565–6576. PMLR, 2021.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Alexander Cosgrove, Christopher D Manning, Christopher Re, Diana Acosta-Navas, Drew Arad Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue WANG, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri S. Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Andrew Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. Holistic evaluation of language models. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=iO4LZibEqW. Featured Certification, Expert Certification.

Yang Liu, Yuanshun Yao, Jean-Francois Ton, Xiaoying Zhang, Ruocheng Guo Hao Cheng, Yegor Klochkov, Muhammad Faaiz Taufiq, and Hang Li. Trustworthy LLMs: a survey and guideline for evaluating large language models' alignment, 2023. Preprint at https://arxiv.org/abs/2308.05374.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Ro{bert}a: A robustly optimized {bert} pretraining approach, 2020. URL https://openreview.net/forum?id=SyxS0T4tvS.

Mark Hugo Lopez, Jens Manuel Krogstad, and Jeffrey S. Passel. Who is hispanic? *Pew Research Center*, 2023. URL https://www.pewresearch.org/short-reads/2023/09/05/who-is-hispanic/.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019. Preprint at https://arxiv.org/abs/1711.05101.

Sandra Martinková, Karolina Stanczak, and Isabelle Augenstein. Measuring gender bias in West Slavic language models. In Jakub Piskorski, Michał Marcińczuk, Preslav Nakov, Maciej Ogrodniczuk, Senja Pollak, Pavel Přibáň, Piotr Rybak, Josef Steinberger, and Roman Yangarber (eds.), *Proceedings of the 9th Workshop on Slavic Natural Language Processing 2023 (SlavicNLP 2023)*, pp. 146–154, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.bsnlp-1.17. URL https://aclanthology.org/2023.bsnlp-1.17/.

Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. SemEval-2018 task 1: Affect in tweets. In Marianna Apidianaki, Saif M. Mohammad, Jonathan May, Ekaterina Shutova, Steven Bethard, and Marine Carpuat (eds.), *Proceedings of the 12th International Workshop on Semantic Evaluation*, pp. 1–17. Association for Computational Linguistics, New Orleans, Louisiana, June 2018. doi: 10.18653/v1/S18-1001. URL https://aclanthology.org/S18-1001.

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

Roberto Navigli, Simone Conia, and Björn Ross. Biases in large language models: Origins, inventory, and discussion. *J. Data and Information Quality*, 15(2), jun 2023. ISSN 1936-1955. doi: 10.1145/3597307. URL https://doi.org/10.1145/3597307.

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

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 53728–53741. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/a85b405ed65c6477a4fe8302b5e06ce7-Paper-Conference.pdf.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. Beyond accuracy: Behavioral testing of NLP models with CheckList. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4902–4912. Association for Computational Linguistics, Online, July 2020. doi: 10.18653/v1/2020.acl-main.442. URL https://aclanthology.org/2020.acl-main.442.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. The risk of racial bias in hate speech detection. In Anna Korhonen, David Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 1668–1678. Association for Computational Linguistics, Florence, Italy, July 2019. doi: 10.18653/v1/P19-1163. URL https://aclanthology.org/P19-1163.

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

Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. The woman worked as a babysitter: On biases in language generation. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3407–3412. Association for Computational Linguistics, Hong Kong, China, November 2019. doi: 10.18653/v1/D19-1339. URL https://aclanthology.org/D19-1339.

Emily Sheng, Kai-Wei Chang, Prem Natarajan, and Nanyun Peng. Societal biases in language generation: Progress and challenges. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 4275–4293. Association for Computational Linguistics, Online, August 2021. doi: 10.18653/v1/2021.acl-long.330. URL https://aclanthology.org/2021.acl-long.330.

Vered Shwartz and Yejin Choi. Do neural language models overcome reporting bias? In Donia Scott, Nuria Bel, and Chengqing Zong (eds.), *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 6863–6870. International Committee on Computational Linguistics, Barcelona, Spain (Online), December 2020. doi: 10.18653/v1/2020.coling-main.605. URL https://aclanthology.org/2020.coling-main.605.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In David Yarowsky, Timothy Baldwin, Anna Korhonen, Karen Livescu, and Steven Bethard (eds.), *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pp. 1631–1642. Association for Computational Linguistics, Seattle, Washington, USA, October 2013. URL https://aclanthology.org/D13-1170.

Jacob-Junqi Tian, Omkar Dige, David Emerson, and Faiza Khan Khattak. Interpretable stereotype identification through reasoning, 2023. Preprint at https://arxiv.org/abs/2308.00071.

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. Llama 2: Open foundation and fine-tuned chat models, 2023. Preprint at https://arxiv.org/abs/2307.09288.

Nikolai Sergeevich Trubetzkoy. *Principles of Phonology*. University of California Press, 1969.

Yixin Wan, George Pu, Jiao Sun, Aparna Garimella, Kai-Wei Chang, and Nanyun Peng. "Kelly is a warm person, Joseph is a role model": Gender biases in LLM-generated reference letters. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 3730–3748. Association for Computational Linguistics, Singapore, December 2023. doi: 10.18653/v1/2023.findings-emnlp.243. URL https://aclanthology.org/2023.findings-emnlp.243.

Linda R Waugh. Marked and unmarked: A choice between unequals in semiotic structure. *Linguistics*, 1982.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22. Curran Associates Inc., Red Hook, NY, USA, 2024. ISBN 9781713871088.

Robert Wolfe and Aylin Caliskan. Markedness in visual semantic ai. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, pp. 1269–1279. Association for Computing Machinery, New York, NY, USA, 2022a. ISBN 9781450393522. doi: 10.1145/3531146.3533183. URL https://doi.org/10.1145/3531146.3533183.

Robert Wolfe and Aylin Caliskan. American == white in multimodal language-and-image ai. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '22, pp. 800–812. Association for Computing Machinery, New York, NY, USA, 2022b. ISBN 9781450392471. doi: 10.1145/3514094.3534136. URL https://doi.org/10.1145/3514094.3534136.

Abdelrahman Zayed, Goncalo Mordido, Ioana Baldini, and Sarath Chandar. Why don't prompt-based fairness metrics correlate? In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9002–9019, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.487. URL https://aclanthology.org/2024.acl-long.487/.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, 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. OPT: Open pre-trained transformer language models, 2022. Preprint at https://arxiv.org/abs/2205.01068.

A Fine-Tuning Hyperparameters

For completeness, we provide the full details of the hyperparameter tuning process used in the fine-tuning experiments. During fine-tuning, early stopping is applied based on validation loss. If no improvement in the loss is observed over a fixed number of steps, then training is stopped. An AdamW optimizer is used with default parameters, except for learning rate (LR) and weight decay (Loshchilov & Hutter, 2019). A hyper-parameter study was performed to select the best early stopping threshold and LR for all models. For fully fine-tuned models, weight decay was also optimized.

The early stopping threshold was varied between five and seven steps. The learning rate (LR) was chosen from {1e-3, 3e-4, 1e-4, 3e-5, 1e-5}, and weight decay, when tuned, was selected from {1e-3, 1e-4, 1e-5, 1e-6}. For larger models, LoRA fine-tuning was applied with the rank parameter 8 on every non-embedding layer.

For RoBERTa 125M and 355M and OPT 125M and 350M, 15 training runs were performed, and the five models with the highest accuracy on the SST5 test set were retained. For the larger models, due to resource constraints, five models in total were trained for each model type. Table 1 summarizes the optimal hyperparameters selected for each model during fine-tuning.

Table 1: Hyperparameters used for model fine-tuning.

Model	Early stop threshold	LR	Weight decay
RoBERTa-125M	7	1e−5	1e−5
RoBERTa-355M	7	1e−5	1e−5
OPT-125M	7	1e−5	1e−5
OPT-350M	7	1e−5	1e−3
OPT-1.3B	5	1e−4	1e−4
OPT-6.7B	5	1e−4	1e−4
OPT-13B	5	1e−4	1e−4
Llama-2-7B	5	1e−4	1e−4
Llama-2-13B	5	1e−4	1e−4
Llama-3-8B	5	1e−4	1e−3
Mistral-7B	5	3e−5	1e−3

B Prompt Templates and Other Details

This section includes the templates used in the prompting approach. Each subsection corresponds to a different template. For CoT prompting, inference batches are limited to size 4 due to higher computational demands, whereas batch sizes of 16 are used in previous settings.

B.1 Zero-Shot Prompt Template

The zero-shot prompt template is displayed below with additional formatting for readability. The component in angled brackets is where each sample to be classified is inserted. The models begin generation at [*LM Generation*].

Text: ⟨Text to classify⟩
Question: Is the sentiment of the text negative, neutral, or positive?
Answer: The sentiment is [*LM Generation*]

B.2 Few-Shot Prompt

Below is the few-shot template. For the few-shot prompt templates, nine labeled examples are prepended to the prompt, following the template style. The models begin generation at [*LM Generation*].

Text: Example 1 from either SST5 or SemEval
Question: What is the sentiment of the text?
Answer: Negative.

...

Text: Example 9 from either SST5 or SemEval
Question: What is the sentiment of the text?
Answer: Positive.

Text: ⟨Text to classify⟩
Question: What is the sentiment of the text?
Answer: [*LM Generation*]

B.3 Zero-Shot CoT Prompt

Zero-shot CoT uses two prompt templates in sequence. In the first step, the model is provided the text to classify and asked about the corresponding sentiment. The traditional "trigger" sentence "Let's think step by step" is used to encourage the model to generate reasoning prior to answering the question. The template appears below.

Text: ⟨Text to classify⟩
Question: Is the sentiment of the text negative, neutral, or positive?
Reasoning: Let's think step by step. [*LM Generation*]

In the second step of zero-shot CoT, the reasoning generation is appended to the first prompt along with the answer completion text displayed in the template below. At this stage, the model is expected to generate an answer to be extracted.

Text: ⟨Text to classify⟩
Question: Is the sentiment of the text negative, neutral, or positive?
Reasoning: Let's think step by step. ⟨Generation from previous step⟩
Answer: Therefore, from negative, neutral, or positive, the sentiment is [*LM Generation*]

C Fine-tuning and Prompting Accuracy

Tables 2 and 3 presents the average classification accuracy and the standard deviation for the fine-tuning and prompting approaches on the Amazon dataset, respectively. Generally, prompt-based classification accuracy is lower than that of fine-tuning.

Table 2: Accuracy statistics on the Amazon dataset for fine-tuning experiments across model types and sizes. Bold numbers indicate the best accuracy achieved within each model family.

Model	Size	Mean Accuracy	Standard Deviation
RoBERTa	125M	**0.635**	0.036
	350M	0.624	0.027
OPT	125M	0.687	0.080
	350M	0.692	0.039
	1.3B	**0.739**	0.020
	6.7B	0.737	0.014
Llama-2	7B	0.513	0.089
	13B	**0.647**	0.006
Llama-3	8B	**0.822**	0.035
Mistral	7B	**0.740**	0.005

Table 3: Model accuracy and standard deviation (in parentheses) on the Amazon dataset for prompting experiments across model types. Bold numbers indicate the best accuracy achieved for each model.

Prompt Type	Zero-shot	Zero-shot CoT	SemEval 9-shot	SST5 9-shot
OPT-6.7B	0.451 (0.002)	–	**0.482** (0.009)	0.433 (0.024)
Llama-2-7B	0.483 (0.002)	0.492 (0.003)	**0.654** (0.037)	0.616 (0.028)
Llama-3-8B	0.600 (0.003)	0.539 (0.001)	0.683 (0.017)	**0.716** (0.024)
Mistral-7B	0.502 (0.003)	0.517 (0.003)	**0.700** (0.045)	0.682 (0.025)

D Gap Differences Across Models

For each of the models, across the different datasets, an FPR gap span is calculated. For a given type of FPR, Negative- or Positive-Sentiment gap spans are computed as the largest difference between any two mean FPR gaps for the groups. This quantifies how large the particular FPR disparities for a given model and dataset are between groups. The larger this span, the greater the difference in Negative- or Positive-Sentiment FPR between groups and the less invariant the model is to overall group substitution. Table 4 displays the FPR gap spans for each model, averaged over the three datasets. In computing the spans, the gap for the White group is part of the span extremes 58% of the time for Negative-Sentiment FPR and 100% of the time for Positive-Sentiment FPR. That is, the gap computed for the White group often constitutes one of the largest gap magnitudes.

From the table, it is clear that the RoBERTa and Llama-2 models have consistently large spans for both types of FPR gap. On the other hand, Llama-3-8B, the most recent model studied, has the smallest average gap spans in both categories. Another recent model, Mistral-7B, demonstrates a small average Negative-Sentiment FPR gap span, suggesting that more recent LLMs may be slightly less affected by issues with the template-based probes. It is interesting to note that the distribution of spans for Positive-Sentiment FPR gaps are more uniformly distributed between models than the Negative-Sentiment counterpart.

Rank	Model	Mean Negative-Sentiment FPR Gap Span	Model	Mean Positive-Sentiment FPR Gap Span
1	Llama-2-13B	0.207	RoBERTa-355M	0.154
2	RoBERTa-355M	0.198	RoBERTa-125M	0.152
3	Llama-2-7B	0.144	OPT-13B	0.144
4	RoBERTa-125M	0.126	Llama-2-13B	0.143
5	OPT-350M	0.118	Mistral-7B	0.141
6	OPT-1.3B	0.104	OPT-125M	0.136
7	OPT-6.7B	0.081	Llama-2-7B	0.133
8	OPT-13B	0.056	OPT-350M	0.132
9	OPT-125M	0.039	OPT-1.3B	0.128
10	Mistral-7B	0.032	OPT-6.7B	**0.089**
11	Llama-3-8B	**0.020**	Llama-3-8B	**0.089**

Table 4: Models ranked by average gap spans across datasets for Negative- and Positive-Sentiment FPR when fine-tuning. For a given type of FPR, gap spans are computed as the largest difference between any two mean FPR gaps across groups. The larger this span, the greater the difference in Negative- or Positive-Sentiment FPR between groups and the less invariant the model is to group substitution.