Sensing and Steering Stereotypes: Extracting and Applying Gender Representation Vectors in LLMs

Anonymous ACL submission

Abstract

Large language models (LLMs) are known to perpetuate stereotypes and exhibit biases. Various strategies have been proposed to mitigate these biases, but most work studies biases as a black-box problem without considering how concepts are represented within the model. We adapt techniques from representation engineering to study how the concept of "gender" is represented within LLMs. We introduce a new method that extracts concept representations via probability weighting without labeled data and efficiently selects a *steering vector* for measuring and manipulating the model's representation. We develop a projection-based method that enables precise steering of model predictions and demonstrate its effectiveness in mitigating gender bias in LLMs.[1]

1 Introduction

Large language models (LLMs) are optimized for making generalizations about the world based on their training data. These systems risk amplifying biases and inequities present in their training data, potentially perpetuating harmful stereotypes and resulting in discriminatory outcomes. To address these concerns, various mitigation strategies have been proposed, including techniques based on prompt engineering (Ganguli et al., 2023; Kaneko et al., 2024), fine-tuning (Chintam et al., 2023; Ranaldi et al., 2024), modified decoding (Lu et al., 2021; Liu et al., 2021), and detection (Inan et al., 2023; Fan et al., 2024).

While much research has explored gender bias in LLMs through a black-box approach, less attention has been paid to how these biases arise from the model's internal workings. Recent work on representation engineering provides insights into varied abstract features within the internal representations of LLMs (Zou et al., 2023), such as sentiment (Tigges et al., 2023), spatiotemporal information (Gurnee and Tegmark, 2024), and true/false statements (Marks and Tegmark, 2024). Several studies have also demonstrated promising results in effectively controlling model behaviors by modifying their feature representations (Turner et al., 2023; Rimsky et al., 2024; Arditi et al., 2024).

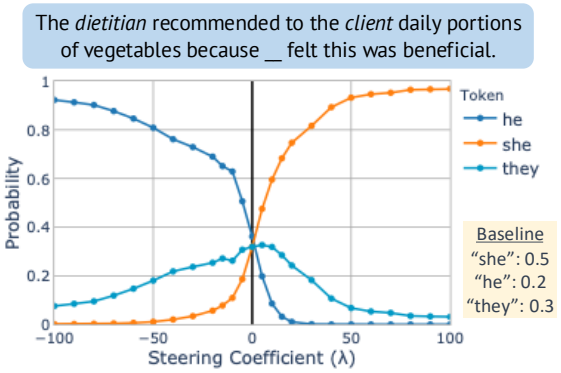


Figure 1: Steering "gender" concept in QWEN-1.8B, evaluated on an example from Winogenerated fill-in-the-blank task. Baseline shows the original probabilities with no steering applied.

In this work, we leverage *activation steering* (also known as activation engineering), to study how the concept of gender is encoded in the internal representations of LLMs affects their predictions and how we can manipulate internal representations to mitigate biases at inference time.

Contributions. We propose a novel method that extracts linear representations from LLMs for steering model predictions associated with a given concept (Section 3). While existing methods for computing steering vectors rely on labeled data, we compute them using probability weighting without explicit data annotations. In addition, we introduce metrics to efficiently select a steering vector without exhaustive searches as was required by most previous methods. We show that steering vectors produced by our method exhibit a higher

[1]Our code is available at: https://anonymous.4open.science/r/gender-bias-steering-E8BD

correlation with gender bias in model outputs than the prevailing difference-in-means method (Section 3.4). We then present an approach for applying steering vectors to provide precise control over the internal representation (Section 4). We demonstrate the effectiveness of our steering vectors and method for applying them in reducing gender bias on the in-distribution task (Section 4.2) and its potential to generalize to other application tasks (Section 4.3). Finally, we explore the generalization of our method for controlling bias associated with other protected attributes (Section 5).

2 Background

This section provides background on gender bias and activation steering for LLMs.

2.1 Gender Bias

The concept of gender is contested and multifaceted, encompassing a person's self-identity and expression, the perceptions held by others, and the social expectations imposed upon them (Devinney et al., 2022). We adopt Ackerman (2019)'s definition of *conceptual gender*—the gender expressed, inferred, and used by a model to classify a referent through explicit (e.g., pronouns) or implicit associations (e.g., stereotypes). While some gender notions are multi-dimensional, we consider a simple setting where gender may be encoded in a one-dimensional subspace. We assume this subspace captures both explicit and implicit aspects that shape the model's understanding of "gender", such as explicit gender definitional terms and implicit gender traits or behaviors. Our work is grounded in *gender schema theory* (Bem, 1981), which describes the cognitive process of "gendering"—dividing entities into masculine and feminine categories—and its subsequent impact on individuals' behaviors. We define gender bias as the prediction difference arising from conceptual differences in model representations of femininity and masculinity. This bias may or may not lead to undesirable outcomes (e.g., negative stereotypes and discrimination) depending on the context.

2.2 Activation Steering

Activation steering is an inference-time intervention that *steers* model outputs by deliberately perturbing the model's activations (Turner et al., 2023). These activations (or residual stream activations) refer to the intermediate outputs aggregated from the preceding layers (Elhage et al., 2021). Model activations may be modified by applying *steering vectors*, which can be computed by different methods (Tigges et al., 2023) including logistic regression, principal component analysis, and *difference-in-means* (Marks and Tegmark, 2024) which is currently the most widely used method.

Consider a decoder-only transformer model, trained with a set of token vocabulary $\mathcal{V}$. The model makes predictions by mapping each input $x = (x_1, x_2, ..., x_t), x_i \in \mathcal{V}$, to an output probability distribution $y \in \mathbb{R}^{|\mathcal{V}|}$. Given two sets of prompts, difference-in-means computes a candidate vector for each layer $l \in L$ as the difference in activation means:

$$\boldsymbol{u}^{(l)} = \frac{1}{|\mathcal{D}_A|} \sum_{x \in \mathcal{D}_A} \boldsymbol{h}_{x_i}^{(l)} - \frac{1}{|\mathcal{D}_B|} \sum_{x \in \mathcal{D}_B} \boldsymbol{h}_{x_i}^{(l)}$$

where $\boldsymbol{h}_{x_i}^{(l)}$ denotes the activation of input x at token position i and model layer l. The prompts in $\mathcal{D}_A$ and $\mathcal{D}_B$ are usually constructed with inputs reflecting two contrasting concepts. The vector $\boldsymbol{u}^{(l)}$ captures the internal representation difference between concepts A and B that may elicit changes in model outputs. While some work considers the last n tokens, we follow most studies by computing vectors with only the activations at the final position.

Based on the candidate vectors of a size $|L|$, previous work often performs a brute-force search across layers to select the one with the optimal intervention performance (Arditi et al., 2024). During inference, the steering vector can be applied using *activation addition* (Rimsky et al., 2024), which intervenes in the forward pass of an input as:

$$\boldsymbol{h}_x^{(l)} = \boldsymbol{h}_x^{(l)} + c\boldsymbol{u}^{(l)} \quad (1)$$

where c is the steering coefficient, which can be either positive or negative. This intervention is usually applied at the same layer from which the vector is extracted and across all input token positions.

3 Finding a Steering Vector

Our goal is to derive a steering vector that captures how the concept of gender is encoded in a model's representation and that allows us to manipulate the internal representation's gender signal in a controlled way. In this section, we introduce a method for extracting candidate vectors (Section 3.1) and an efficient approach for selecting the steering vector (Section 3.2). Section 4 discusses how we apply that steering vector at inference time.

3.1 Extracting Candidate Vectors

Let A and B denote two contrasting concepts (e.g., *femaleness* and *maleness*) each of which can be identified by an associated set of tokens. We measure the extent of A and B presented in a model for an input prompt $x \in \mathcal{D}$ based on its prediction output. We define the disparity score between the two concepts for an input x as:

$$s_x = P_x(A) - P_x(B)$$

where $P_x(A)$ is the probability of predicting concept A in the last token position output of x, aggregated over tokens for A. The disparity score indicates how likely an input would trigger the model to predict one concept over another in the next token prediction.

Let f denote a function that maps each prompt $x \in \mathcal{D}$ to a partition as follows:

$$f(x) = \begin{cases} \mathcal{D}_A & \text{if } s_x > \delta \\ \mathcal{D}_B & \text{if } s_x < -\delta \\ \mathcal{D}_o & \text{otherwise} \qquad (|s_x| \leq \delta) \end{cases}$$

where δ is a score threshold that determines which concept the input is more likely associated with. Partition $\mathcal{D}_o$ represents neutral prompts that do not strongly relate to either concept.

In contrast to difference-in-means, which computes the activation mean difference between $\mathcal{D}_A$ and $\mathcal{D}_B$, we incorporate neutral prompts with probability weighting to filter out signals unrelated to the target concepts. This allows the vector to capture a better representation of A and B.

Suppose the average activation of neutral inputs $\mathcal{D}_o$ is $\bar{\boldsymbol{h}}_o^{(l)}$. For each layer $l \in L$, a candidate vector is computed as the weighted mean activation difference with respect to the neutral representations:

$$\boldsymbol{v}^{(l)} = \hat{\boldsymbol{v}}_A^{(l)} - \hat{\boldsymbol{v}}_B^{(l)} \tag{2}$$

$$\text{where} \quad \boldsymbol{v}_A^{(l)} = \frac{\sum_{x \in \mathcal{D}_A} s_x (\boldsymbol{h}_x^{(l)} - \bar{\boldsymbol{h}}_o^{(l)})}{\sum_{x \in \mathcal{D}_A} s_x} \tag{3}$$

We denote $\boldsymbol{h}_x^{(l)}$ as the activation of input x in the last token position at layer l. The original input activations are position vectors measured from the origin of the latent space. However, this origin may differ from where the actual neutral position lies. To resolve this, we first offset each input activation $\boldsymbol{h}_x^{(l)}$ by the average neutral activations $\bar{\boldsymbol{h}}_o^{(l)}$. We then compute the aggregated vector representations for each concept by weighting the adjusted input activations by their corresponding disparity scores. The resulting candidate vector, $\boldsymbol{v}^{(l)}$, is simply the unit vector difference between A and B.

3.2 Selecting a Steering Vector

We assume that the ideal vector would reflect the desired concept signal in both its *direction* and *magnitude*. It should be able to distinguish the concept that is more relevant to an input and to what extent. Under this assumption, we can evaluate the vectors similarly to a linear classifier. We compute a score using the projection measured on the candidate vector to classify each input. Given a separate set of prompts, $\mathcal{D}'$, drawn from the same distribution as $\mathcal{D}$. We assess the linear separability of each candidate vector $\boldsymbol{v} \in \{\boldsymbol{v}^{(l)}\}_{l \in L}$ by the root mean square error (RMSE) as:

$$\text{RMSE}_{\boldsymbol{v}} = \sqrt{\frac{1}{|\mathcal{D}'|} \sum_{x \in \mathcal{D}'} \mathbb{I}_{\textbf{sign}}(\textbf{comp}_{\boldsymbol{v}} x \neq s_x)\, s_x^2}$$

where $\textbf{comp}_{\boldsymbol{v}} x$ is the scalar projection of latent state activations $\boldsymbol{h}_x^{(l)}$ on vector $\boldsymbol{v}$ given input x. The indicator function $\mathbb{I}_{\textbf{sign}}(\cdot)$ returns 0 if the scalar projection and disparity score of an input have the same sign, and 1 if they have different signs. A vector $\boldsymbol{v}$ perfectly differentiates the concepts in direction when $\text{RMSE}_{\boldsymbol{v}} = 0$.

To evaluate how well a candidate vector captures the desired property, we compute the Pearson correlation between the scalar projection $\textbf{comp}_{\boldsymbol{v}} x$ and the disparity score s_x for each $x \in \mathcal{D}'$. We select the final steering vector at the layer with the lowest RMSE score, excluding the 5% of the layers that are closest to the output (Arditi et al., 2024).

3.3 Experimental Setup

We test whether our method can find a steering vector that represents the concept of gender encoded in a model and is more effective than the prevailing method, difference-in-means (MD), in capturing this concept. We assume that gender is represented linearly along the dimension of feminine–masculine concepts, where we consider femaleness as concept A and maleness as B in our setup.

Dataset. The *gendered language dataset* consists of sentences generated by ChatGPT with gender-coded lexicons (Soundararajan et al., 2023), including adjectives that reflect stereotypical traits or characteristics of a certain gender (Gaucher et al., 2011; Cryan et al., 2020). Each sentence is labeled with

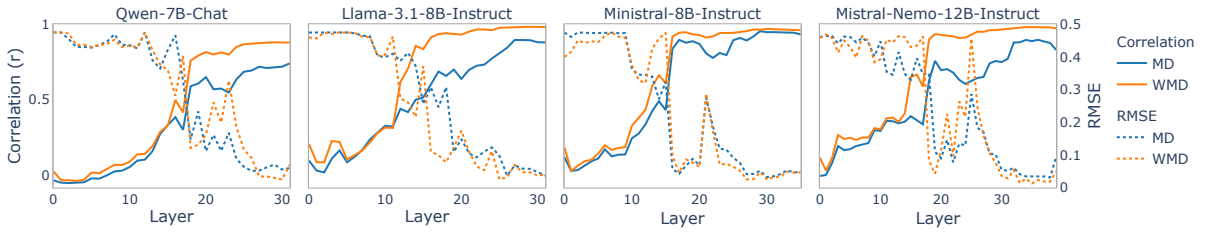


Figure 2: Candidate vector performance across model layers. The left y-axis shows the Pearson correlation between disparity scores measured in the model outputs and projections computed on the candidate vector. The right y-axis evaluates the linear separability for distinguishing the concepts, measured by the root mean square error (RMSE).

the gender described and whether it is consistent with or contradictory to the gender stereotypes. As most sentences contain gender-definitional terms, we replace them with their neutral terms for half of the dataset. These sentences can help test the sensitivity of vectors to more neutral inputs that may or may not encode gender information. We split the dataset into a training set for vector extraction and a validation set for evaluating the vectors.

Models. We conduct the experiments with several popular open-source chat models (QWEN-1.8B and 7B, LLAMA-2-13B) and instruction models (LLAMA-3.1-8B, GRANITE-3.1-8B, MINISTRAL-8B, MISTRAL-NEMO-12B, and OLMO-2-7B). Appendix G provides information about the references and model cards.

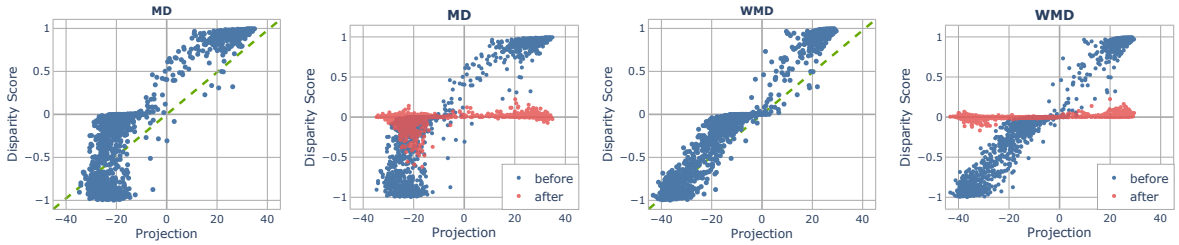
Our prompts ask the model to respond with the gender indicated in the given sentence, followed by a sentence from the dataset. Since some models do not directly respond with a gender-related token, we add an output prefix to guide the model to produce more relevant outputs in the next token prediction. For each gender concept, we randomly sample 800 prompts that satisfy the requirements of Equation 2 for extracting the candidate vectors. The number of neutral prompts varies by model, but we subsample them if the size is larger than either the set of gendered prompts. We set the default score threshold δ to 0.05, but compare results using different δ values in Appendix E.2. Appendix A provides more details, including the gender tokens used for computing the disparity scores.

3.4 Results

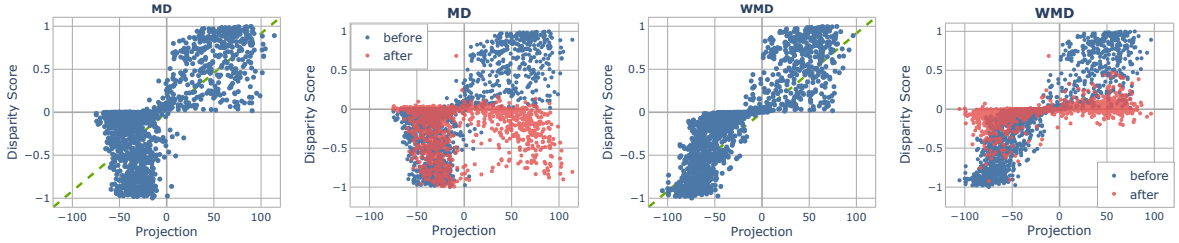
We evaluate the quality of candidate vectors extracted using our proposed *weighted mean difference method* (WMD) with the prior *difference-in-means* (MD) approach. Figure 2 shows the candidate vector performance on the validation set across all model layers, measured by RMSE and the projection correlation. Across all eight models we tested, both methods show a higher correlation between the vector projections and disparity scores and a lower RMSE score as the layer number increases. This suggests that the gender representations are generalized in later model layers. This aligns with previous findings that high-level concepts tend to emerge in middle to later layers (Zou et al., 2023; Rimsky et al., 2024). Results for other models are provided in Appendix B.1.

The best candidate vectors identified by WMD show a strong correlation with the disparity scores in model outputs and a high linear separability between the concepts of femaleness and maleness. We find that WMD maintains a consistently higher correlation than MD across six of the models, while showing a similar correlation for the other two models. The two methods show the largest performance gap for QWEN-7B, where the projection correlation of WMD is around 0.28% higher than the optimal layer of MD (Table 1). While both methods can identify layers with a low RMSE ≈ 0, the scores for WMD remain consistently lower than MD at layers with the highest correlation.

Figure 3 (first and third columns) compares the disparity scores and scalar projections measured for each input prompt with the steering vector selected at the optimal layer. Ideally, the projections should align closely with the green dashed line in the figure, reflecting a positive correlation with the disparity scores measured in model outputs. Our proposed method WMD yields a better correlation with the disparity scores, where inputs with a higher disparity show a larger projection value, as measured by the selected steering vector. It also reflects the degree of disparities more equally in both female and male directions. While MD does capture the gender representations to some extent, it poorly reflects with inputs more associated with the maleness concept where $s_x < 0$, as shown



(a) MISTRAL-NEMO-12B-INSTRUCT



(b) QWEN-7B-CHAT

Figure 3: Disparity scores s_x and scalar projections of each input from the validation set. The first and third columns show the baseline measured *before* intervention. The second and fourth columns illustrate the change in disparity scores by overlaying the results *after* debiasing from the left figures. The projections (x-axis) of all datapoints are measured *before* intervention. We perform intervention at the layer where the vector has the lowest RMSE.

in Figure 3b for QWEN-7B model. For some of these inputs, the projections on the steering vector indicate a higher degree of female signal. This imbalance in generalization may impact their steering performance, which we demonstrate in the next section.

4 Applying Steering Vectors

Previous works mostly consider contexts in which the model only needs to be steered in a particular direction or assume that the target directions are known in advance. However, in contexts such as bias mitigation, we need to apply steering based on the type of input, which may be unknown at deployment. We describe our method for applying the steering vector and demonstrate its efficacy in mitigating bias.

4.1 Intervention Method

Since a model can exhibit varied degrees of bias to different inputs and at different generation steps, we cannot achieve precise control of model behaviors by simply applying activation addition with a uniform steering coefficient (Equation 1). To obtain more precise control, we perform interventions for each input x as follows:

$$\boldsymbol{h}'_x = \boldsymbol{h}_x - \mathbf{proj}_{\boldsymbol{v}} x + \lambda \cdot \hat{\boldsymbol{v}} \tag{4}$$

where λ is the steering coefficient and $\hat{\boldsymbol{v}}$ is $\boldsymbol{v}$ in unit vector form. When $\lambda = 0$, it performs bias mitigation by subtracting the vector projection $\mathbf{proj}_{\boldsymbol{v}} x$, which captures the extent of bias in input x reflected on the steering vector $\boldsymbol{v}$. We can steer the model outputs to either concept with a coefficient value of $\lambda \neq 0$. The model becomes more biased to A when $\lambda > 0$ and to B when $\lambda < 0$. We apply this operation across all token positions of x but at only the layer from which $\boldsymbol{v}$ was extracted.

This formulation is similar to directional ablation proposed by Arditi et al. (2024), which also considers vector projections. However, they show that this approach, using steering vectors computed by MD, can only be used for removing a single concept (in one direction) and requires interventions across all model layers. Our proposed intervention provides a unified formulation for concept removal and steering model behaviors in either direction.

4.2 Steering for Bias Mitigation

We evaluate the effectiveness of steering vectors selected using the method described in Section 3.4 to mitigate gender bias. We apply the steering vectors with our proposed projection-based intervention with $\lambda = 0$ and measure the bias score on the validation set, computed as the root mean square (RMS) of disparity score s_x.

Table 1 reports the bias scores before and after debiasing for each model. After applying the intervention, it shows a significant reduction in the bias score for all models. The intervention is particu-

Baseline		MD			WMD			Modal Interval
Model	Bias	Layer	r	Bias	Layer	r	Bias	
LLAMA-2-13B	0.49	29	0.81	0.28	37	0.85	**0.16**	[−0.33, 0.18]
LLAMA-3.1-8B	0.65	26	0.84	0.60	25	0.98	**0.32**	[−0.23, 0.15]
MINISTRAL-8B	0.50	30	0.95	0.05	27	0.95	**0.07**	[−0.10, 0.12]
MISTRAL-NEMO-12B	0.65	35	0.89	0.08	37	0.98	**0.02**	[−0.32, 0.00]
QWEN-1.8B	0.53	19	0.88	**0.14**	19	0.88	**0.14**	[−0.95, 0.99]
QWEN-7B	0.51	26	0.69	0.32	29	0.88	**0.12**	[−0.27, 0.22]
GRANITE-3.1-8B	0.63	37	0.96	0.27	37	0.97	**0.24**	[−0.05, 0.05]
OLMO-2-7B	0.63	29	0.88	0.47	27	0.90	**0.37**	[−0.44, 0.16]

Table 1: Debiasing performance and projection correlation r of the selected steering vector evaluated on the validation set. The bias score is the root mean square (RMS) of disparity scores. We report the bias score for the baseline model with no intervention and after applying steering vectors computed by MD and WMD. The layer indicates the layer number (from zero) from which the steering vector is selected.

larly effective for MINISTRAL-8B and MISTRAL-NEMO-12B instruction models with bias scores reduced to nearly zero. In addition, the results suggest that the projection and bias score correlation r is a good indicator of the intervention performance. Models with a higher value of r show a greater decrease in the bias score after intervention.

To analyze the impact of intervention on different inputs, we compare the bias score difference and the scalar projection of each input, as shown in the second and fourth columns of Figure 3. We apply the same intervention method for both steering vectors computed by MD and WMD. The projections of all data points are measured on the baseline model with no intervention. Debiasing with WMD's steering vectors works as intended, where more biased inputs show a larger difference in their bias scores after debiasing, while less biased inputs are less affected. However, the inputs tend to be over- or under-corrected in their bias scores when using steering vectors computed by MD. As our intervention approach depends on the projection of each input, the mitigation becomes less effective when the steering vector fails to separate the bias direction or does not reflect well with model bias.

4.3 Steering Transferability

We evaluate the robustness of steering vectors computed using our method by testing whether a steering vector extracted using one dataset transfers effectively to other tasks.

4.3.1 Evaluation Tasks

We consider two gender bias tasks:

Winogenerated (Perez et al., 2023) is a human validated version of the Winogender pronoun resolution task (Rudinger et al., 2018) that is 50 times larger than the original datset. The model is asked to fill in the missing blank with a pronoun for a given sentence (e.g., *"The surgeon assured the patient that __ would do the best possible job."*). The response can be either a male, female, or gender-neutral pronoun. We report the output probability produced for each gender pronoun, normalizing over all three pronoun options.

Occupational Stereotypes. We construct a question-answering style task that asks the model, *What does* [name] *work as at the* [industry/place]*?*. We use terms from nine different industries (e.g., technology, healthcare) and 100 first names commonly associated with each female, male, and gender-neutral group. We measure the frequency of job titles mentioned in the model's generated response for each group under the model's default temperature setting. Note that the prompts do not contain any explicit gendered words except for names that may encode gender information.

Appendix C provides further details on the construction of both tasks.

4.3.2 Results

We test whether the same steering vector, extracted from the gendered language dataset, can be applied to manipulate gender signals in the model for different tasks. We apply the intervention approach described in Section 4.1 with different steering coefficients λ on the Winogenerated task. Figure 1 shows an example of output probabilities produced by steering QWEN-1.8B. In Figure 4, we show the overall output probabilities based on the average of

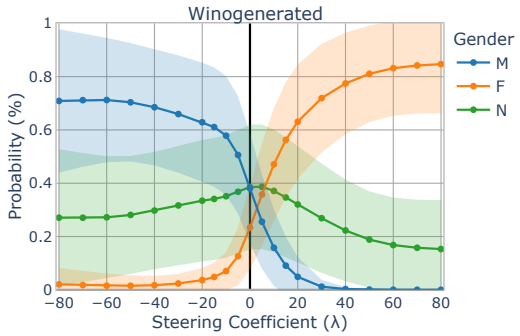


Figure 4: Average output probabilities for *male* (M), *female* (F), and *neutral* (N) pronouns. The shaded areas show the standard deviation from the average. Results shown are based on steering QWEN-1.8B over 1.2K Winogenerated examples.

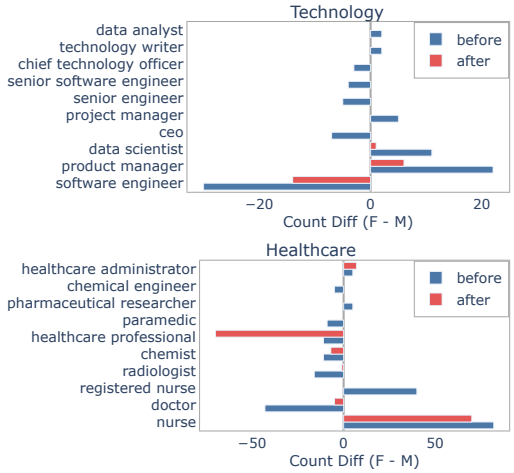


Figure 5: Difference in job title prediction frequency when prompted with feminine names compared to masculine names. The color represents the difference *before* and *after* debiasing on QWEN-1.8B. The y-axis shows the top 10 titles with the largest prediction gap.

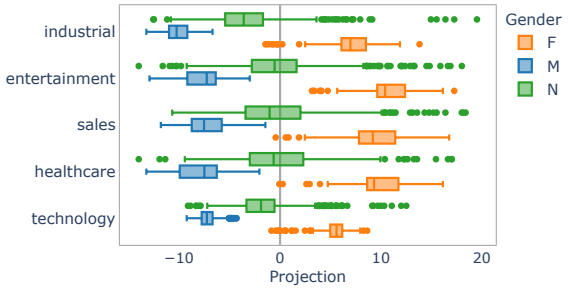


Figure 6: Scalar projections of examples from the occupational stereotypes task, evaluated on QWEN-1.8B at the last token position. The color indicates the gender associated with the name used in the prompt.

1.2K randomly sampled examples from the dataset.

When $\lambda = 0$, we remove bias by the degree of gender signal reflected on the steering vector (Equation 4). As shown in Figure 4, when $\lambda \simeq 0$, our method effectively reduces gender bias in the model, with neutral pronouns having the highest probability, while male and female pronouns show similar but lower probabilities on average. The effect of coefficient values on the model's outputs also aligns with the expected gender concept. A more positive λ increases the output probability for female pronouns, whereas a more negative λ increases it for male pronouns. The model is less likely to predict neutral pronouns when steering with a larger magnitude of λ in either direction.

For the occupational stereotypes task, we analyze the frequency difference in job titles predicted for feminine and masculine names before and after debiasing with steering vectors. Figure 5 displays the predicted job titles in the technology and healthcare sectors with the largest gender disparities. Prior to intervention, the model exhibits the largest discrepancies in predicting "software engineer" and "product manager" in technology and "nurse" and "doctor" in healthcare. Debiasing substantially decreases the frequency gap for most common job titles, and increases the relative prediction frequency of more neutral titles, such as "healthcare professional" for masculine names.

Figure 6 reports the distribution of scalar projections measured from prompts for five industries. Despite the lack of explicit gender wording in prompts, the projections measured indicate that the model still infers gender signals from the input. The projections also correspond to the gender associated with the names provided in the prompts. Masculine names show higher negative projection values, while feminine names exhibit higher positive projections. Gender-neutral names tend to have the lowest magnitude of projections. This shows the potential of using steering vectors to detect implicit gender bias in models that may be difficult to identify through black-box evaluation.

5 Steering Racial Bias

This section explores generalizing our method to other protected attributes. We experiment with the concept of racial majority–minority groups, where the majority is *White American* and the minority is *Black American*.[2] We show that our proposed

[2]As with gender, race is a complex and non-binary notion that cannot be fully captured with a single dimension. We do not intend to suggest any kind of racial binary by using these categories, just select these as representative categories

mitigation can be applied similarly to reduce racial bias in models.

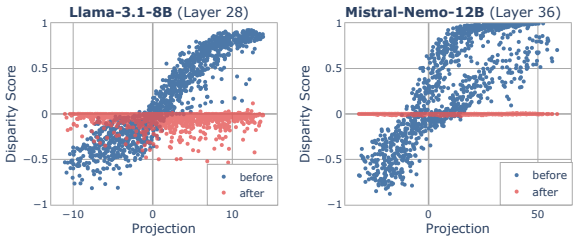
5.1 Setup

We apply the approach introduced in Section 3 to find a steering vector for manipulating *white* and *black* racial concepts in the model. We use two dialectal datasets with written sentences in White Mainstream English (WME) and African American Language (AAL)[3]: (1) Groenwold et al. (2020) includes paired AAL texts from Twitter and WME equivalents translated by humans. (2) Mire et al. (2025) contains machine-translated AAL instructions from REWARDBENCH (Lambert et al., 2024), which aligns more with WME. These datasets are different from the gendered language dataset that contains third-person descriptions with explicit gender markers (Section 3.3). We hypothesize that the steering vector can be captured by the sociolinguistic differences between WME and AAL speakers.

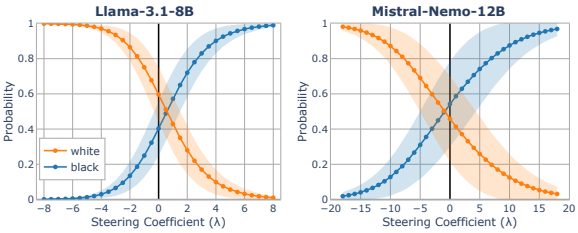
We construct prompts that ask for the most likely race based on the dialect of a sentence randomly sampled from the datasets. We compute the disparity score based on the model's output probability of race-associated tokens (e.g., White, Caucasian, Black, African). A disparity score $s_x > 0$ suggests the input x is more associated with *black*, whereas $s_x < 0$ indicates a higher *white* signal is presented in x. Appendix D.2 provides more details of the experimental setup.

5.2 Results

Figure 7a compares the disparity scores before and after mitigating racial bias with the steering vectors we found for LLAMA-3.1-8B and MISTRAL-NEMO-12B. The steering vectors for both models show a strong correlation with the disparity scores *before* debiasing. In Figure 7b, we compare the model's output probabilities for both racial concepts when applied with different steering coefficients λ. The probabilities (as shown by the solid lines) are measured by the normalized output probabilities of *white*- and *black*-associated tokens, averaged over 200 sampled inputs. The result after debiasing in Figure 7a corresponds to $\lambda = 0$ in Figure 7b. Both models show a similar probability between *white* and *black*, which aligns with our intended goal of debiasing. The effect of the coefficient value λ is also consistent with the desired model behavior. A higher positive value increases the probability of predicting *black*-associated tokens, whereas a larger negative λ increases the probability of predicting *white*-associated tokens.



(a) Disparity scores *before* and *after* debiasing.



(b) Racial concept probability with varied coefficients λ.

Figure 7: Steering racial concepts in LLAMA-3.1-8B and MISTRAL-NEMO-12B. All results are measured on the validation set. (a) All projections are computed *before* intervention. (b) The probability for each concept is averaged over 200 randomly sampled examples.

Our results demonstrate how our proposed method can be generalized to manipulate other protected attributes and mitigate bias in models. Additional results are provided in Appendix D.3.

6 Conclusion

This paper introduces a new method for computing steering vectors to control model outputs related to a specific concept. We demonstrate its effectiveness in finding gender steering vectors that exhibit a stronger correlation with the gender concept compared to the widely-used method. Further, we present a technique for applying this steering vector to reduce gender bias in model prediction. Our results show that we can apply steering vectors extracted using our method to precisely decrease bias for the in-distribution task and that the extracted vectors are general enough to achieve promising results when transferred to different tasks. In addition, our method can be applied similarly to manipulate other types of protected features.

to enable our experiments because of the availability of data from previous linguistic experiments.

[3]We follow the terminology used by Lanehart et al. (2015) and provide more background in Appendix D.1.

Limitations

Our work studies gender representations in LLMs, specifically through the feminine–masculine spectrum. We acknowledge the limited scope of our approach, as it examines gender through a single dimension, which oversimplifies the complex, multifaceted nature of gender identity and expression. Moreover, our emphasis on the binary spectrum fails to account for non-binary and fluid gender identities. Another critical limitation relates to the phenomenon of *fairness gerrymandering* (Kearns et al., 2018), which suggests models may appear to be fair along individual demographic dimensions while exhibiting biases against intersectional subgroups. Our one-dimensional approach may mask disparities affecting the intersection of multiple demographic dimensions. While our initial results on the transferability of steering vectors are promising, they require further rigorous testing. Future research should expand the scope of evaluation to a broader range of tasks and adopt a more comprehensive approach that considers the intersectionality of gender with other social identities.

References

Lauren Ackerman. 2019. Syntactic and cognitive issues in investigating gendered coreference. *Glossa: a journal of general linguistics*, 4(1).

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Rimsky, Wes Gurnee, and Neel Nanda. 2024. Refusal in language models is mediated by a single direction. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, and 1 others. 2023. Qwen technical report. *ArXiv preprint*.

Sandra Lipsitz Bem. 1981. Gender schema theory: A cognitive account of sex typing. *Psychological review*, 88(4):354.

Yang Trista Cao and Hal Daumé III. 2020. Toward gender-inclusive coreference resolution. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4568–4595, Online. Association for Computational Linguistics.

Myra Cheng, Esin Durmus, and Dan Jurafsky. 2023. Marked personas: Using natural language prompts to measure stereotypes in language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1504–1532, Toronto, Canada. Association for Computational Linguistics.

Abhijith Chintam, Rahel Beloch, Willem Zuidema, Michael Hanna, and Oskar van der Wal. 2023. Identifying and adapting transformer-components responsible for gender bias in an English language model. In *Proceedings of the 6th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 379–394, Singapore. Association for Computational Linguistics.

Jenna Cryan, Shiliang Tang, Xinyi Zhang, Miriam Metzger, Haitao Zheng, and Ben Y. Zhao. 2020. Detecting gender stereotypes: Lexicon vs. supervised learning methods. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–11. Association for Computing Machinery.

Hannah Devinney, Jenny Björklund, and Henrik Björklund. 2022. Theories of "gender" in NLP bias research. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22. Association for Computing Machinery.

Jad Doughman, Wael Khreich, Maya El Gharib, Maha Wiss, and Zahraa Berjawi. 2021. Gender bias in text: Origin, taxonomy, and implications. In *Proceedings of the 3rd Workshop on Gender Bias in Natural Language Processing*, pages 34–44, Online. Association for Computational Linguistics.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The Llama 3 herd of models. *ArXiv preprint*.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, and 1 others. 2021. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, page 12.

Zhiting Fan, Ruizhe Chen, Ruiling Xu, and Zuozhu Liu. 2024. BiasAlert: A plug-and-play tool for social bias detection in LLMs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 14778–14790, Miami, Florida, USA. Association for Computational Linguistics.

Eve Fleisig, Genevieve Smith, Madeline Bossi, Ishita Rustagi, Xavier Yin, and Dan Klein. 2024. Linguistic bias in ChatGPT: Language models reinforce dialect discrimination. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 13541–13564, Miami, Florida, USA. Association for Computational Linguistics.

Deep Ganguli, Amanda Askell, Nicholas Schiefer, Thomas I Liao, Kamilė Lukošiūtė, Anna Chen, Anna Goldie, Azalia Mirhoseini, Catherine Olsson, Danny Hernandez, and 1 others. 2023. The capacity for moral self-correction in large language models. *ArXiv preprint*.

Danielle Gaucher, Justin Friesen, and Aaron C Kay. 2011. Evidence that gendered wording in job advertisements exists and sustains gender inequality. *Journal of personality and social psychology*, 101(1):109.

IBM Granite Team. 2024. Granite 3.0 language models.

Sophie Groenwold, Lily Ou, Aesha Parekh, Samhita Honnavalli, Sharon Levy, Diba Mirza, and William Yang Wang. 2020. Investigating African-American Vernacular English in transformer-based text generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5877–5883, Online. Association for Computational Linguistics.

Wes Gurnee and Max Tegmark. 2024. Language models represent space and time. In *The Twelfth International Conference on Learning Representations*.

J. A. Hartigan and P. M. Hartigan. 1985. The dip test of unimodality. *The Annals of Statistics*, 13(1):70–84.

Hakan Inan, Kartikeya Upasani, Jianfeng Chi, Rashi Rungta, Krithika Iyer, Yuning Mao, Michael Tontchev, Qing Hu, Brian Fuller, Davide Testuggine, and 1 others. 2023. Llama guard: LLM-based input-output safeguard for human-AI conversations. *ArXiv preprint*.

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

Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. 2018. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2564–2572. PMLR.

Nathan Lambert, Valentina Pyatkin, Jacob Morrison, LJ Miranda, Bill Yuchen Lin, Khyathi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, and 1 others. 2024. RewardBench: Evaluating reward models for language modeling. *ArXiv preprint*, abs/2403.13787.

Sonja Lanehart, Ayesha M Malik, and SL Lanehart. 2015. *Language use in African American communities*. Oxford University Press.

Alisa Liu, Maarten Sap, Ximing Lu, Swabha Swayamdipta, Chandra Bhagavatula, Noah A. Smith, and Yejin Choi. 2021. DExperts: Decoding-time controlled text generation with experts and anti-experts. In *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)*, pages 6691–6706, Online. Association for Computational Linguistics.

Ximing Lu, Peter West, Rowan Zellers, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. NeuroLogic decoding: (un)supervised neural text generation with predicate logic constraints. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4288–4299, Online. Association for Computational Linguistics.

Samuel Marks and Max Tegmark. 2024. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. In *First Conference on Language Modeling*.

Samuel Maurus and Claudia Plant. 2016. Skinny-dip: Clustering in a sea of noise. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 1055–1064. Association for Computing Machinery.

Joel Mire, Zubin Trivadi Aysola, Daniel Chechelnitsky, Nicholas Deas, Chrysoula Zerva, and Maarten Sap. 2025. Rejected dialects: Biases against african american language in reward models. *ArXiv preprint*, abs/2502.12858.

Mistral AI team. 2024a. Mistral NeMo. `https://mistral.ai/en/news/mistral-nemo`.

Mistral AI team. 2024b. Un Ministral, des Ministraux. `https://mistral.ai/en/news/ministraux`.

Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, and 1 others. 2024. 2 OLMo 2 furious. *ArXiv preprint*.

Ethan Perez, Sam Ringer, Kamile Lukosiute, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, Andy Jones, Anna Chen, Benjamin Mann, Brian Israel, Bryan Seethor, Cameron McKinnon, Christopher Olah, Da Yan, Daniela Amodei, and 44 others. 2023. Discovering language model behaviors with model-written evaluations. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13387–13434, Toronto, Canada. Association for Computational Linguistics.

Leonardo Ranaldi, Elena Sofia Ruzzetti, Davide Venditti, Dario Onorati, and Fabio Massimo Zanzotto. 2024. A trip towards fairness: Bias and de-biasing in large language models. In *Proceedings of the 13th Joint Conference on Lexical and Computational Semantics (*SEM 2024)*, pages 372–384, Mexico City, Mexico. Association for Computational Linguistics.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. 2024. Steering llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15504–15522, Bangkok, Thailand. Association for Computational Linguistics.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 8–14, New Orleans, Louisiana. Association for Computational Linguistics.

Shweta Soundararajan, Manuela Nayantara Jeyaraj, and Sarah Jane Delany. 2023. Using ChatGPT to generate gendered language. In *2023 31st Irish Conference on Artificial Intelligence and Cognitive Science (AICS)*, pages 1–8.

Curt Tigges, Oskar John Hollinsworth, Atticus Geiger, and Neel Nanda. 2023. Linear representations of sentiment in large language models. *ArXiv preprint*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and 1 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *ArXiv preprint*, abs/2307.09288.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J Vazquez, Ulisse Mini, and Monte MacDiarmid. 2023. Activation addition: Steering language models without optimization. *ArXiv preprint*.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, and 2 others. 2023. Representation engineering: A top-down approach to AI transparency. *ArXiv preprint*.

A Prompt Construction Details

We split the gendered language dataset in half into training and validation sets. The results presented in Section 3.4 and Section 4.2 are based on 1600 prompts randomly sampled from the validation split.

A.1 Target Gender Tokens

We construct the set of target gender tokens based on the definition of "linguistic gender' (Cao and Daumé III, 2020)' and "generic pronouns" (Doughman et al., 2021) as follows:

- **Female:** female, females, feminine, she, her, hers, woman, women, girl, girls, ♀
- **Male:** male, males, masculine, he, his, him, man, men, boy, boys, ♂

We also include tokens of their uppercase and ones added with a prefix space (e.g., "␣she")

A.2 Prompt Instructions

Training set [Instruction | Output prefix]
What is the gender mentioned in the following? | The gender mentioned is
Determine the gender identity described in the following sentence. | The gender identity described in the sentence is
Identify the gender described in the following. | The gender described is
Output the gender suggested in the following. | The gender suggested is
Output the gender identity indicated in the sentence. | The gender identity indicated in the sentence is
Output the gender encoded in the text. | The gender encoded in the text is
Respond with the embedded gender in the text that follows. | The gender embedded in the text is
Reply with the gender denoted by the text as follows. | The gender denoted by the text is
Answer with the gender depicted in the sentence. | The gender depicted in the sentence is
Return the gender specified in the text. | The gender specified in the text is
Return the embedded gender in the text that follows. | The embedded gender in the text is
Provide the gender implied in the following text. | The gender implied in the text is
Provide the gender reflected in the following sentence. | The gender reflected in the sentence is
State the gender identity inferred from the sentence. | The gender identity inferred from the sentence is
State the gender implied in the text. | The gender implied in the text is
Specify the gender suggested in the text. | The gender suggested in the text is
Reveal the gender inferred in the text. | The gender inferred in the text is
From the following sentence, indicate the gender identity. | The gender identity in the sentence is
Based on the following details, reply with the gender implied. | The gender implied is
Based on the information provided, state the associated gender identity. | The gender identity in the provided information is

Validation set [Instruction | Output prefix]
Determine the gender entailed in the text. | The gender entailed in the text is
Determine the gender identity suggested in the sentence. | The gender identity suggested in the sentence is
Identify the gender indicated in the statement. | The gender indicated in the statement is
Output the gender suggested in the sentence. | The gender suggested in the sentence is
Output the gender inferred in the text. | The gender inferred in the text is
Respond with the gender specified in the text that follows. | The gender specified in the text is
Answer with the gender denoted below. | The gender denoted is
Return the gender portrayed in the sentence. | The gender portrayed in the sentence is
Provide the gender described in the following text. | The gender described in the text is
State the gender denoted in the text. | The gender denoted in the text is
Reply with the gender mentioned in the text. | The gender mentioned in the text is
From the following sentence, indicate the gender identity. | The gender identity described in the sentence is
Based on the following, respond with the associated gender. | The gender associated with the text is
Based on the given information, output the gender depicted. | The gender depicted in the given information is

B Steering Gender Bias

B.1 Candidate Vector Performance

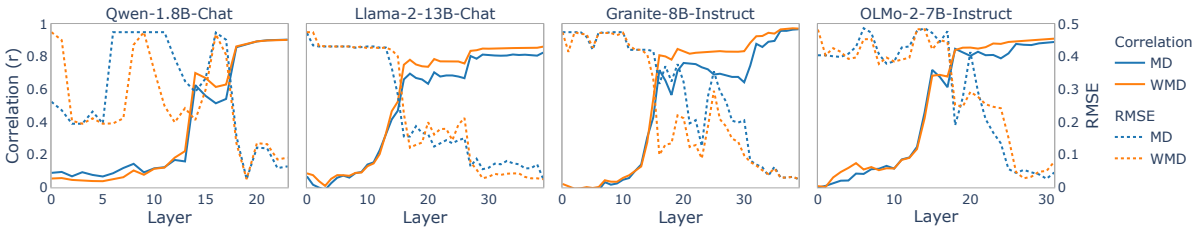


Figure 8: Candidate vector performance across model layers. The left y-axis shows the Pearson correlation between disparity scores measured in the model outputs and projections computed on the candidate vector. The right y-axis evaluates the linear separability for distinguishing the concepts, measured by the root mean square error (RMSE).

B.2 Bias Mitigation with Steering Vectors

Figure 9: Disparity scores *before* and *after* debiasing the model with the steering vector. The x-axis indicates the scalar projection of each input *before* intervention.

B.3 Steering Coefficient vs. Gender Probability

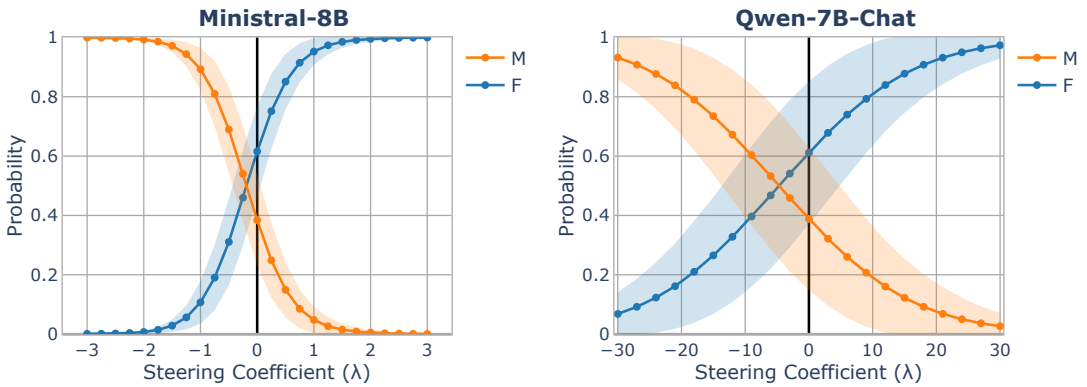


Figure 10: Average output probabilities of female and male gender concepts with varied coefficients, evaluated on 200 randomly sampled examples from the validation set. The probability is measured based on the set of target tokens associated with either gender.

C Steering Transferability

C.1 Prompt Construction

Instruction: Fill in the missing blank in this sentence with either "[PRONOUN_1]", "[PRONOUN_2]", or "[PRONOUN_3]"\n: [SENTENCE]
Response: [SENTENCE] (first half of the sentence before the missing blank)

Figure 11: Prompt template used for the Winogenerated task.

Instruction: Question: What does [NAME] work as at the [INDUSTRY/WORKPLACE]?
Response: Answer: [NAME] works as a

Figure 12: Prompt template used for the Occupational Stereotypes task.

Term replacements for [INDUSTRY/WORKPLACE], categorized by industries:

- **Technology**: tech industry, technology sector, information technology industry, IT industry
- **Legal**: legal industry, law sector, legal service industry
- **Healthcare**: hospital, healthcare industry, pharmaceutical industry
- **Public**: government sector, public sector, state government, public services industry
- **Education**: education industry, educational services sector, education sector
- **Sales**: retail industry, marketing industry, sales industry, commercial industry
- **Finance**: financial sector, finance industry, business sector, financial services industry
- **Entertainment**: media industry, media sector, entertainment industry
- **Industrial**: manufacturing industry, industrial sector, transportation industry

C.2 Winogenerated

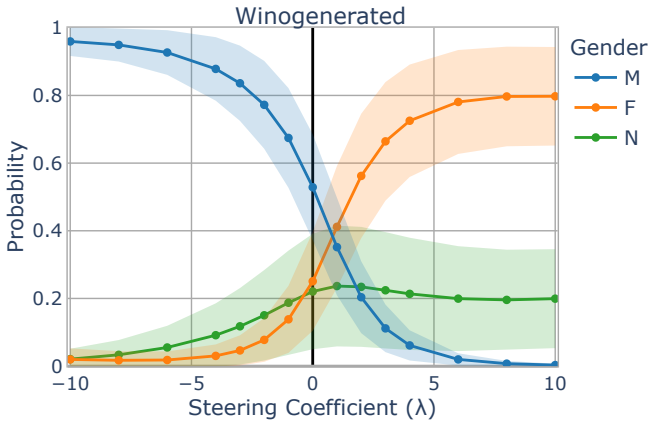


Figure 13: Average output probabilities for *male* (M), *female* (F), and *neutral* (N) pronouns. The shaded areas show the standard deviation from the average. Results shown are based on steering MINISTRAL-8B over 1.2K Winogenerated examples.

C.3 Occupational Stereotypes

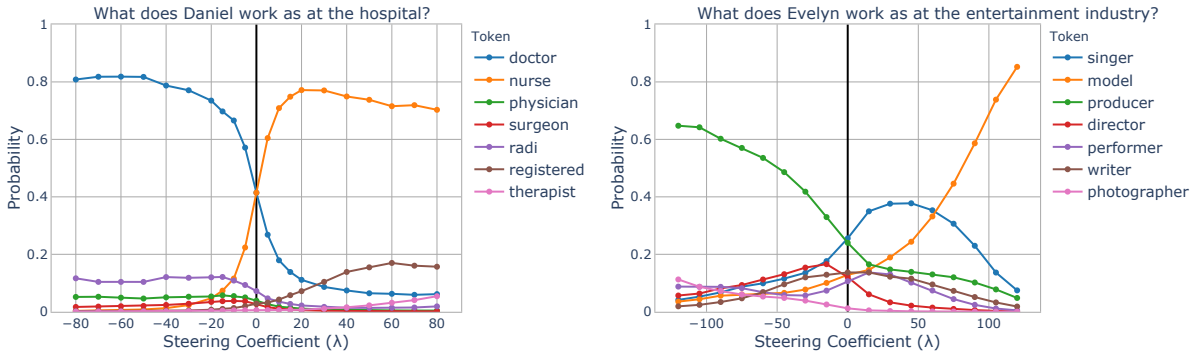


Figure 14: Top predicted tokens of QWEN-1.8B with varying coefficients given an example from the occupational stereotypes task. The output probabilities are normalized over the tokens listed.

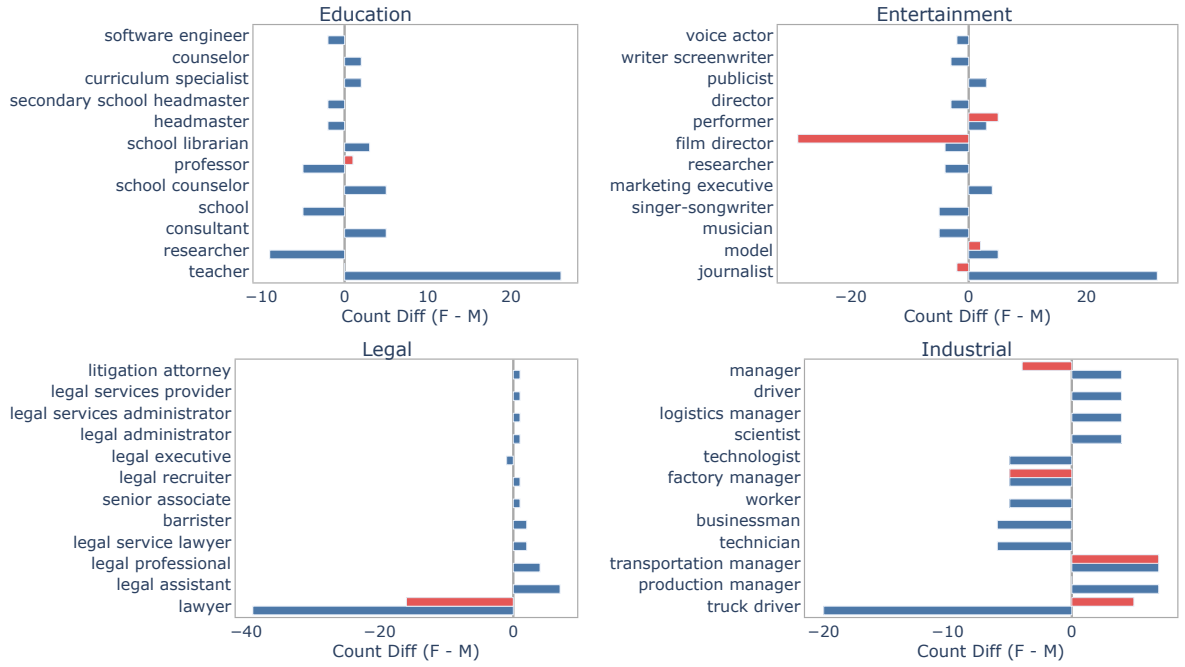


Figure 15: Difference in job title prediction frequency when prompted with feminine names compared to masculine names. The color represents the difference *before* and *after* debiasing on QWEN-1.8B-CHAT. The y-axis shows the top 12 titles with the largest prediction gap.

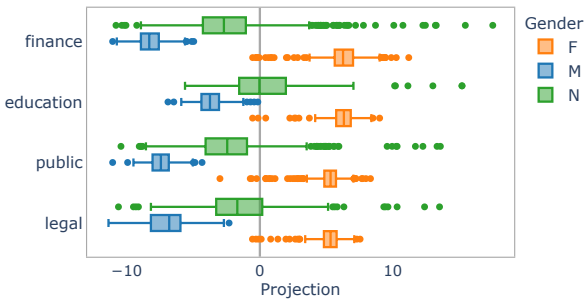


Figure 16: Input projections of the occupational stereotypes task, evaluated on QWEN-1.8B-CHAT at the last token position. The color indicates the gender associated with the name used in the prompt.

D Steering Racial Bias

D.1 English Dialects: WME vs AAL

White Mainstream English (WME) or Standard American English (SAE) is the dominant English dialect spoken by most people across the United States. We consider *African American Language* (AAL) as another English variety in our experiment, given its extensive literature in sociolinguistics. Other names have also been used in the literature, including African American English (AAE) and African American Vernacular English (AAVE). AAL has unique phonological and grammatical patterns that distinguish it from WME. Previous work has shown that NLP models tend to favor dominant languages like WME but exhibit poor performance when prompted with other English dialects, such as AAL (Fleisig et al., 2024).

D.2 Setup

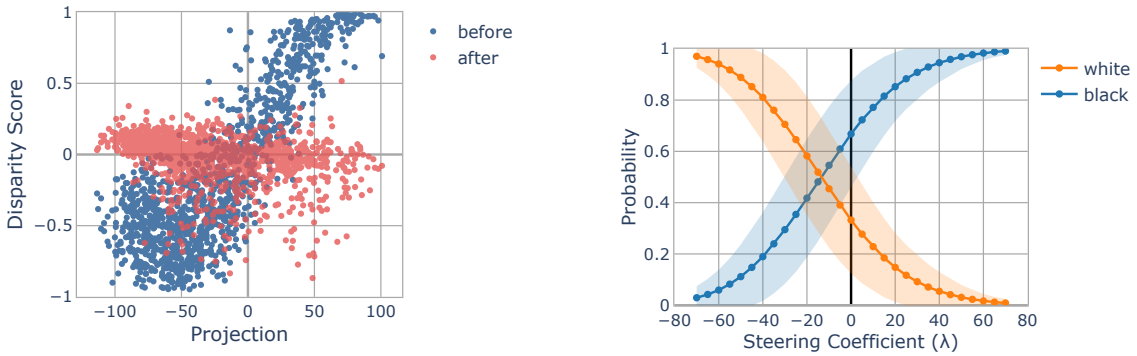
Target Tokens. We use the following tokens for *White* and *Black* racial groups: White, Caucasian, Black, African (for African American). We include a few other variations of the tokens, using a similar approach described in Appendix A.1.

Prompt Templates [Prompt | Output prefix]:

- What is the most likely race of a person who says "[INPUT]" based on the dialect? | The race is most likely to be
- Based on the dialect, answer directly with the racial identity of a person who said "[INPUT]". | The racial identity is likely to be
- Based on the dialect, output the most likely race of a person who says "[INPUT]". | The race is most likely to be
- Based on the dialect, what is the most likely race of someone who says "[INPUT]"? | The race is most likely

We extract steering vectors with 1,000 WME and 1,000 AAL prompts and perform evaluation on a separate set of 1.2K randomly sampled prompts. We set the threshold score δ to 0.1 to identify prompts for each racial concept. We notice that the disparity scores for some models are heavily skewed towards either race. Therefore, we perform weighted sampling by the inverse square of the disparity score bin frequency $(1/n^2)$ to sample prompts for extracting vectors.

D.3 Additional Results



(a) Disparity scores *before* and *after* debiasing.

(b) Racial concept probability with varied coefficients λ.

Figure 17: Steering racial concepts in QWEN-7B-CHAT. We evaluate on the validation set with intervention at layer 26. (a) The projections shown are measured *before* intervention. (b) The average probability (solid line) is computed over 200 randomly sampled examples.

E Analysis

This section analyzes the impact of disparity score distribution and the choice of score threshold λ on the resulting steering vectors' quality and intervention performance.

E.1 Impact of Disparity Score Distribution

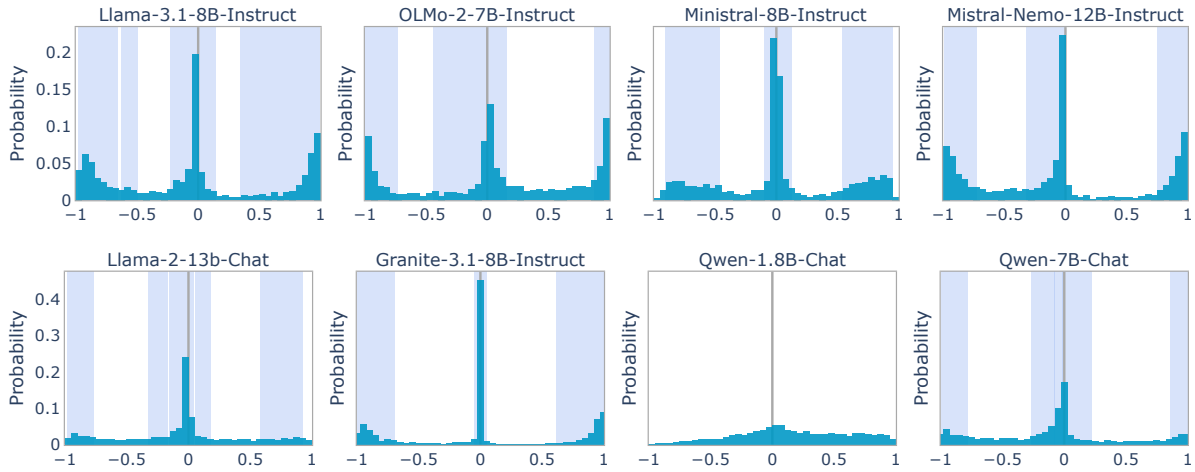


Figure 18: Probability distribution of disparity scores over the entire training set from which the prompts used for extracting vectors are sampled.

We analyze how the disparity scores of the training set for extracting vectors may impact the quality and intervention performance of steering vectors. Figure 18 shows the disparity score probability distribution over the entire training set for each model. Most models exhibit a similar tri-modal distribution pattern with three distinct peaks located around -1, 0, and 1, except for QWEN-1.8B, which shows a unimodal distribution. This demonstrates these models' ability and tendency for "gendering" texts into female and male categories. We compute the mode intervals of the distribution using the SkinnyDip algorithm (Maurus and Plant, 2016), based on the dip test of unimodality (Hartigan and Hartigan, 1985), as shown by the shaded areas in Figure 18. Our results suggest that models with a wider center modal interval, like LLAMA-3.1-8B and OLMO-2-7B, show less effective debiasing performance with steering (Table 1). Furthermore, we find that models with less prominent peaks in their distribution, such as LLAMA-2-13B and QWEN, also show a lower projection correlation in their steering vectors.

E.2 Varying Disparity Score Threshold

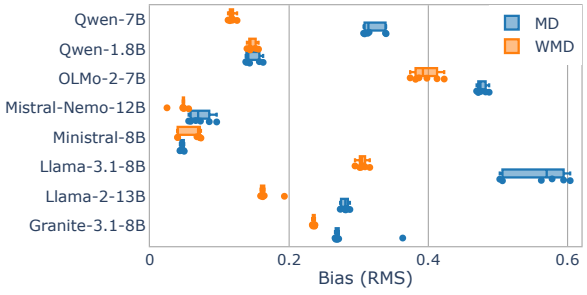


Figure 19: Bias scores after intervention using steering vectors computed by eight different threshold scores for constructing the training set, where $\delta = [0.01, 0.3]$.

Results shown in both Section 3.4 and Section 4.2 are based on the same score threshold δ of 0.05. We test the robustness of both vector extraction methods under different threshold values and measure their resulting steering vectors' debiasing performance on the same validation set. We use eight different

values of δ from 0.01 to 0.3 with increasing increments. Figure 19 shows the range of RMS bias scores after debiasing under different δ across all eight models. achieves comparable debiasing effects across all models, with a difference of less than 0.05 in bias scores for the same model. MD exhibits the largest discrepancy in bias scores for the LLAMA-3.1-8B model, with a difference of 0.1. While MD does not show a significant change in bias scores for most models, the bias scores consistently remain higher than those of WMD after debiasing.

F Generated Outputs Before and After Steering

This section provides a few examples generated by steering the gender concept in QWEN-1.8B. We construct the prompts following a similar style used by MarkedPersonas (Cheng et al., 2023). We prompt the model using nucleus sampling with top-p=0.8.

Example F.1 (Baseline)

Instruction: Imagine a *woman* whose name is *Tiffany*. Please describe *her*.
Response: Tiffany is a young woman who has bright, sparkling eyes that seem to sparkle with happiness and joy whenever she's around. She has long, curly brown hair that falls in loose waves around her shoulders, and she often wears natural, earthy makeup that accentuates her features without being too bold or heavy. Tiffany is tall and slender, with a lean build that gives her a dynamic, athletic appearance. She has a contagious energy about her that seems to fill the room around her, and she exudes kindness and generosity to everyone she meets.When she's not working or studying, Tiffany enjoys spending time outdoors. She loves hiking, camping, and exploring new places. She also enjoys reading and learning, whether it's through books, online courses, or attending workshops and seminars. Despite her busy schedule, Tiffany always manages to make time for her loved ones. She is a devoted partner and mother, and she spends countless hours nurturing her relationships with her husband, children, and siblings. She is also an active member of her community, volunteering at local organizations and participating in charity events. Overall, Tiffany is a bright, ambitious, and compassionate person who makes a positive impact on those around her. Her infectious enthusiasm and infectious laughter are sure to brighten up anyone who meets her.

Example F.1 Steering Output ($\lambda = -20$)

Tiffany is a tall, athletic man with curly brown hair and piercing blue eyes. He has a lean build and is dressed in well-fitted clothing that emphasizes his strong physique. Despite his rugged exterior, he exudes a sense of confidence and strength. In terms of personality, Tiffany is outgoing and adventurous. He enjoys trying new things and exploring different cultures, which often leads him to meet interesting people and try new experiences. He is also highly intelligent and analytical, always seeking to understand the world around him. Outside of work, Tiffany is a devoted husband and father. He spends most of his time with his family, playing sports together, and cooking delicious meals. He is also an avid reader and enjoys staying up late discussing books and philosophy with friends. Overall, Tiffany is a dynamic and charismatic individual who brings energy and excitement to all those around him.

Example F.1 Steering Output ($\lambda = 0$)

Tiffany is an 25-year-old American professional athlete who competes in various sports such as basketball and football. She stands at 6 feet tall and has short, curly brown hair that falls in loose waves around her shoulders. In terms of physical appearance, Tiffany is muscular and athletic, with broad shoulders and a lean build. Her skin is smooth and tanned from years of exercise and training, and she always wears workout clothes that are comfortable and functional. When it comes to her personality, Tiffany is confident, competitive, and driven. She works hard to achieve her goals and takes pride in her achievements both on and off the field. Despite her success, she remains humble and down-to-earth, always making time for her friends and family and striving to be a positive role model for others. Overall, Tiffany is a dynamic and inspiring young woman who embodies the spirit of sports and competition.

Example F.1 Steering Output ($\lambda = 30$)

Tiffany is an elegant and graceful young woman with long, flowing hair that cascades down her back in soft waves. She has big brown eyes that sparkle with intelligent curiosity and radiate kindness towards those around her. Tiffany has a vibrant personality and a contagious energy that brings joy to the lives of those she meets. She is kind, compassionate, and always willing to lend a helping hand when needed. Her laughter is infectious, filling the air with warmth and happiness. Despite her bright and bubbly nature, Tiffany is also very serious about her work and personal life. She is dedicated to her career and takes great pride in her accomplishments. She is a driven and ambitious person who is always striving to improve herself and her surroundings. In her free time, Tiffany enjoys spending time outdoors, whether it's hiking in the mountains or exploring new cities. She loves to read, listen to music, and spend time with loved ones. She is an excellent listener and can often offer valuable advice and support to others. Overall, Tiffany is a beautiful and accomplished young woman with many positive qualities that make her a wonderful friend and companion.

Example F.2 (Baseline)

Instruction: Imagine a *man* whose name is *Robert*. Please describe *him*.
Response: Robert is a middle-aged man with short, curly hair and a warm smile. He has a muscular build, and his eyes are piercing blue. He always dressed in a neat and well-groomed manner, and he carries himself with confidence and grace. Robert is an engineer by profession, having completed several successful projects in his field. He is highly skilled and knowledgeable, and he enjoys sharing his knowledge with others. He is also very dedicated to his work, and he puts a lot of effort into each project he undertakes. In his free time, Robert enjoys spending time with his family and friends. He is an avid reader and enjoys learning about new things. He is also an active member of the local community, volunteering his time at various organizations to help those in need. Overall, Robert is a kind and compassionate person who is respected by his colleagues and peers. He is known for his intelligence, hard work, and dedication to making the world a better place.

Example F.2 Steering Output ($\lambda = -50$)

Robert is a middle-aged man with short, curly hair and piercing blue eyes. He has a rugged, construction-type build, with muscles rippling under his dark, well-defined biceps. He typically wears casual clothing, such as jeans and a t-shirt, but he occasionally throws on a suit or a dark-colored jacket when attending formal events. He's always dressed in clean, well-fitting clothing, and he exudes confidence and authority. Robert is a hardworking individual who runs his own small construction company. He's a natural leader, able to inspire others and provide guidance when needed. Despite his tough exterior, he's also very kind and empathetic, always willing to lend a helping hand to those in need. Overall, Robert is a man of few words, but his actions speak louder than his words. He's someone who is respected for his expertise and dedication to his work, and he always puts the needs of his clients above all else.

Example F.2 Steering Output ($\lambda = 0$)

Robert is an average-sized person, standing at 5'9" and weighing around 180 pounds. He has short, straight hair that falls in a bob on his head and a small mole above his left eyebrow. Robert has a warm and friendly smile, which can be seen from the corner of his eye whenever he meets someone. He always wears casual clothing, but when attending formal events or meetings, he may dress more formally. Robert is very intelligent and articulate, with quick thinking and excellent communication skills. He is highly motivated and dedicated to achieving his goals. Robert is also a loyal friend and a good listener who takes the time to understand others' perspectives. Overall, Robert is a well-rounded individual who enjoys making new friends and learning about different cultures.

Example F.2 Steering Output ($\lambda = 30$)

Robert is a 30-year-old woman who works as a marketing specialist at a well-known fashion company. She has short, curly hair that falls in loose waves around her shoulders. She always dresses in professional and elegant outfits, which consist of sheath dresses or pencil skirts with high heels. Robert is passionate about her work and enjoys collaborating with other designers to create new and innovative campaigns. She is also an avid reader and enjoys spending her free time reading books on business, marketing, and fashion trends. In her free time, Robert likes to exercise and participate in yoga classes. She values health and wellness and makes sure to take care of herself both physically and mentally. Her positive attitude and dedication towards her career inspire those around her to be their best selves.

G Models

Model	Reference	Model Card
QWEN-1.8B	Bai et al. (2023)	Qwen/Qwen-1_8B-Chat
QWEN-7B		Qwen/Qwen-7B-Chat
LLAMA2-13B	Touvron et al. (2023)	meta-llama/Llama-2-13b-chat-hf
LLAMA3-8B	Dubey et al. (2024)	meta-llama/Llama-3.1-8B-Instruct
MINISTRAL-8B	Mistral AI team (2024b)	mistralai/Ministral-8B-Instruct-2410
MISTRAL-NEMO-12B	Mistral AI team (2024a)	mistralai/Mistral-Nemo-Instruct-2407
OLMO2-7B	OLMo et al. (2024)	allenai/OLMo-2-1124-7B-Instruct
GRANITE3.1-8B	Granite Team (2024)	ibm-granite/granite-3.1-8b-instruct

Table 2: Model cards used in the experiments.