Local Contrastive Editing of Gender Stereotypes

Marlene Lutz
University of Mannheim
marlene.lutz@uni-mannheim.de

Rochelle Choenni
University of Amsterdam
r.m.v.k.choenni@uva.nl

Markus Strohmaier
University of Mannheim, GESIS, CSH Vienna
markus.strohmaier@uni-mannheim.de

Anne Lauscher
University of Hamburg
anne.lauscher@uni-hamburg.de

Abstract

Stereotypical bias encoded in language models (LMs) poses a threat to safe language technology, yet our understanding of how bias manifests in the parameters of LMs remains incomplete. We introduce *local contrastive editing* that enables the localization and editing of a subset of weights in a target model *in relation* to a reference model. We deploy this approach to identify and modify subsets of weights that are associated with gender stereotypes in LMs. Through a series of experiments, we demonstrate that local contrastive editing can precisely localize and control a small subset (<0.5%) of weights that encode gender bias. Our work (i) advances our understanding of how stereotypical biases can manifest in the parameter space of LMs and (ii) opens up new avenues for developing parameter-efficient strategies for controlling model properties in a contrastive manner.

1 Introduction

Stereotypical bias encoded in language models (LMs) can adversely affect the fairness and inclusivity of language technology applications for all users (Blodgett et al., 2020; Choenni et al., 2021; Ma et al., 2023). While considerable efforts have been devoted to measuring (Nadeem et al., 2020; Caliskan et al., 2017) and mitigating (Lauscher et al., 2021) such biases, our understanding of where they manifest in the parameter space of LMs remains limited. Precisely pinpointing biases within the parameters of LMs could enable the development of more targeted and informed bias mitigation strategies. While current research (Ma et al., 2023; Meissner et al., 2022; Hauzenberger et al., 2023) has explored identifying and modifying subcomponents of LMs for bias mitigation, we still lack a thorough understanding of the precise manifestation of biases such as stereotypes in specific model weights.

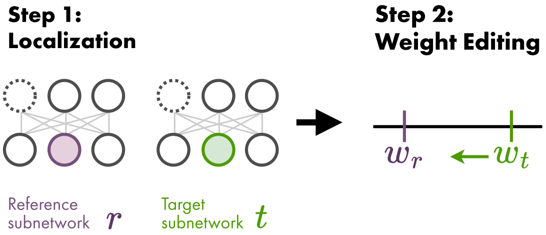


Figure 1: **Local contrastive editing.** In step 1, we localize weights within a target model that encode a certain property. In Step 2, we modify these selected weights relative to a reference model.

Research Goal Consequently, this work aims to (i) localize individual weights that drive stereotypical gender bias in LMs and (ii) modify these weights to steer and mitigate the bias.

Approach We present *local contrastive editing*, a two-step approach that enables the localization and modification of a subset of weights within a target model, relative to a reference model, to control bias (see Figure 1). In step 1, we pinpoint individual weights that encode gender stereotypes via unstructured pruning (Chen et al., 2020). In step 2, we deploy various *local* editing strategies such as weight interpolation or pruning to adjust the identified weights in relation to a reference model.

Results and Contributions We demonstrate the feasibility of local constrastive editing for controlling stereotypical gender bias through a series of experiments. Using our approach, we are able to identify subsets of weights that drive stereotypical bias in LMs. We find that our local editing strategies can flexibly steer gender bias while at the same time retaining the functionality of the model. We provide experimental evidence that most strategies enable a smooth and controllable transition of bias between networks and empirically find that a small subset of weights (<0.5%) is already sufficient to

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 21474–21493
November 12-16, 2024 ©2024 Association for Computational Linguistics

successfully modify and, finally, mitigate the measurable bias.

2 Related Work

Gender Bias Gender bias is present throughout the entire NLP pipeline, from training data (Leavy et al., 2020) to model representations (Bolukbasi et al., 2016; Gonen and Goldberg, 2019), and predictions (Dong et al., 2024). Consequently, much effort has been put towards locating (Joniak and Aizawa, 2022; Chintam et al., 2023) and mitigating gender bias at various stages (Sun et al., 2019; Lauscher et al., 2021; Hauzenberger et al., 2023). We study how gender bias manifests in LMs by detecting and editing a minimal set of relevant parameters to control it.

Knowledge Localization Pruning methods have been used to uncover subnetworks, i.e. subsets of model parameters (Frankle and Carbin, 2018) isolating *task-specific* (Nooralahzadeh and Sennrich, 2023), *domain-specific* (Hendy et al., 2022) or *language-specific* (Wang et al., 2020; Choenni et al., 2023a,b; Nooralahzadeh and Sennrich, 2023) information. In this paper, we use pruning to find subnetworks that contain stereotypical gender bias. Previous research (Vig et al., 2020; Chintam et al., 2023) suggests that stereotypical gender bias is concentrated in specific substructures of a network such as attention heads (Chintam et al., 2023; Ma et al., 2023; Vig et al., 2020) or neurons (Vig et al., 2020). We aim to pinpoint the individual weights responsible for encoding gender bias within a network via unstructured pruning (Chen et al., 2020).

Model Editing Pretrained LMs serve as backbone for many downstream applications, requiring them to be tailored to specific needs. However, the growing size of language models has made traditional fine-tuning costly, leading to increased interest in alternative refinement methods that avoid gradient updates (Yao et al., 2023). One such line of research focuses on efficient model weight editing strategies (Ilharco et al., 2022a,b; Gueta et al., 2023), using mathematical operations on weight vectors composed from the full model to modify information. In this paper, we take a more fine-grained approach to model editing, and instead focus on editing only a subset of weights that we identify as being of relevance for encoding gender stereotypical biases beforehand.

3 Local Contrastive Editing

We localize and adjust specific weights in a target model that are responsible for encoding properties such as stereotypical bias. To achieve this, we use several contrastive strategies based on comparing a *target network* with a *reference network* that differ in a property of interest. We refer to this group of techniques as *contrastive weight editing*.

Formally, let $f(x, \theta)$ be the output of a network with parameters $\theta \in \mathbb{R}^d$ for an example input x. Given a target network $f_t(\cdot, \theta_t)$ and a reference network $f_r(\cdot, \theta_r)$ of the same architecture, with $\theta_t, \theta_r \in \mathbb{R}^d$, we aim to edit θ_t w.r.t. θ_r to modify a property of interest p in f_t while maintaining performance on the original fine-tuning task.

3.1 Localization

In the first step, we investigate how a specific property p manifests in the parameter space of a model and try to localize the individual weights associated with it. To this end, we use unstructured magnitude pruning (Chen et al., 2020) and discover subnetworks in a target and a reference network that are linked to the encoding of p. We define a subnetwork for a network $f(\cdot, \theta)$ as $f(\cdot, m \odot \theta)$, where $\odot$ is the element-wise product and $m \in \{0, 1\}^d$ is a binary pruning mask that sets some parameters in θ to 0. By comparing the target and reference subnetworks, we aim to identify subsets of weights that are related to the encoding of p. We note that subnetworks extracted from different parent models can differ in two aspects: (1) their *pruning masks* may set different parameters to 0; and (2) their parameters θ may have different *values*.

We explore both aspects separately by first selecting the corresponding subsets of weights and then using them to modify the target network. To this end, we define a *localization mask* as the outcome of a particular localization strategy, which indicates which weights will be edited in the subsequent step. Formally, we define such a mask for a given index set $\mathcal{I} \subseteq \{1, \ldots, d\}$ as $b := b(\mathcal{I}) \in \{0, 1\}^d$ via its elements, such that $b_i = \mathbb{1}\{i \in \mathcal{I}\}$. A value of 1 at index i indicates that the corresponding weight is selected for editing. We propose the following two strategies to compute such localization masks.

Mask-based Localization Given a target subnetwork $f_t(\cdot, m_t \odot \theta_t)$ and a reference subnetwork $f_r(\cdot, m_r \odot \theta_r)$, we select those weights that are present in only one of the subnetworks, indicated

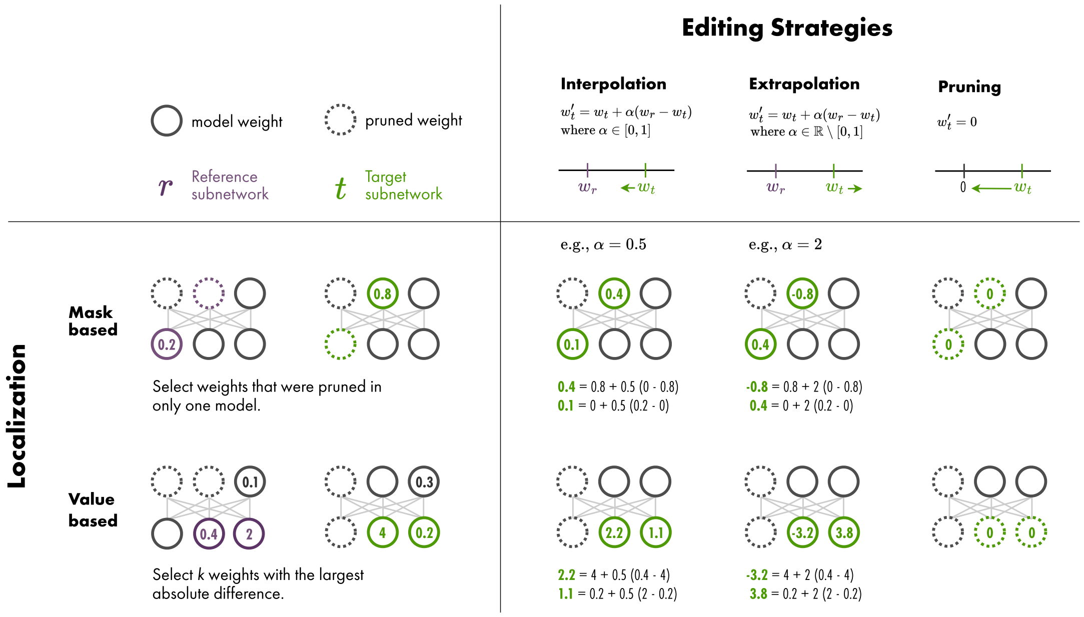


Figure 2: **Overview of localization and editing strategies.** We show value-based and masked-based localization together with our editing strategies (inter- and extrapolation, pruning) on exemplary target and reference networks.

by the pruning masks m_t and m_r . Formally, we compute the localization mask b as:

$$b = m_t \odot m_r \,. \tag{1}$$

We hypothesize that precisely because weights are pruned in one network, but not the other, they encode information relevant to the property p.

Value-based Localization Given a target subnetwork $f_t(\cdot, m_t \odot \theta_t)$ and a reference subnetwork $f_r(\cdot, m_r \odot \theta_r)$, we select a subset of top-k weights that are present in both subnetworks, but differ the most with regard to their values. Let I^k_{top} be the index set containing the indices of the k largest absolute weight differences $|(m_r \odot \theta_r) - (m_t \odot \theta_t)|$. Then we define the localization mask b as:

$$b_i = \begin{cases} 1, & \text{if } i \in I^k_{top} \\ 0, & \text{otherwise}\,. \end{cases} \tag{2}$$

We hypothesize that the weights with the largest absolute difference most strongly steer the networks towards opposing directions with respect to p.

3.2 Contrastive Editing Strategies

After identifying subsets of weights potentially associated with the property of interest p, we use these weights to modify the target network. We explore different types of edits and evaluate their effectiveness. In the following, we assume that we are given a target subnetwork $f_t(\cdot, m_t \odot \theta_t)$, a reference subnetwork $f_r(\cdot, m_r \odot \theta_r)$ and a localization mask b that indicates which weights should be edited. The goal of each of the local editing strategies is to create a new target network $f'_t(\cdot, \theta'_t)$ that is modified with respect to the reference network.

Weight Interpolation (IP) Inspired by recent work on model merging (Ilharco et al., 2022a; Wortsman et al., 2022; Yadav et al., 2024), we propose linear weight interpolation that moves the localized weights of the target network closer to those of its reference or even adopts them completely ($\alpha = 1$):

$$\theta'_t = \theta_t + \alpha((\theta_r - \theta_t) \odot b), \alpha \in [0,1]\,. \tag{3}$$

Note, that linear interpolation can also be used with mask-based localization by assuming that pruned weights have a value of 0.

Weight Extrapolation (EP) Similar to interpolation, we propose linear weight extrapolation to move the localized weights of the target either towards or away from those of the reference network:

$$\theta'_t = \theta_t + \alpha((\theta_r - \theta_t) \odot b), \alpha \in \mathbb{R} \setminus [0,1]\,. \tag{4}$$

Allowing for weighting factors $\alpha \in \mathbb{R} \setminus [0,1]$ enables flexible modifications, including e.g. the removal of a property from a network.

Pruning (PR) Pruning is motivated by the assumption that the localized weights encode a property that can be eliminated by removing precisely those weights:

$$\theta'_t = \theta_t - (\theta_t \odot b)\,. \tag{5}$$

Mask Switch (SW) Our final editing strategy is only applicable for mask-based localization and relies on the impact of weights being present ("turned on") or pruned ("turned off"). We apply the subnetwork mask of the reference model to the target model, resulting in pruning additional weights from the target model. Weights that were initially pruned in the target model during the localization step, but reinstated via the reference subnetwork mask, are restored to their values before pruning.

$$\theta_t' = \theta_t \odot m_r \, . \tag{6}$$

4 Experimental Setup

We showcase the effectiveness of local contrastive editing in one of the, arguably, most established experimental environments for testing bias modification methods from the literature: stereotypical binary gender bias encoded in the BERT [1] model (Devlin et al., 2019). BERT is a widely used transformer model with 12 attention heads and 110 million parameters in total.

4.1 Reference and Target Models

To localize and edit the encoding of stereotypical bias using contrastive strategies, we begin by establishing appropriate target and reference models. For obtaining a thorough understanding of the expected effects, we start from an "extreme" setup in which we intentially bias two types of models to be either *stereotypical* or *anti-stereotypical* concerning specific gender associations. This is accomplished by fine-tuning BERT on subsets of the English Wikipedia [2] that we pre-processed to exhibit gendered associations using the well-established Counterfactual Data Augmentation (Zhao et al., 2018). We describe the process in more detail in the following sections.

4.1.1 Bias Specification

We investigate binary stereotypical gender bias in terms of stereotypical gender associations that manifest in written text. We make use of an explicit bias specification $B = (T_1, T_2, A_1, A_2)$ (Caliskan et al., 2017; Lauscher et al., 2020) that consists of two sets of target words T_1, T_2 that describe demographic groups between which we expect a bias w.r.t. two sets of attributes A_1, A_2. We choose terms in T_1 to represent the female gender (e.g. *woman*) and terms in T_2 to describe the male gender (e.g. *man*). We then build pairs of corresponding terms $(t, t') \subset T_1 \times T_2$ (e.g. (*aunt*, *uncle*)). Further, we designate terms in A_1 to be stereotypically associated with T_1 (e.g. *child-care*) and A_2 to contain attributes that are stereotypically associated with T_2 (e.g. *programming*). The full specification can be found in appendix A and was adopted from Barikeri et al. (2021). Note, that we do not claim our list of target and attribute words to be complete, we rather aim for a small and precise specification that demonstrates the feasibility of our approach.

4.1.2 Counterfactual Data Augmentation

Starting from the bias specification in 4.1.1, we create two datasets that we consider to be stereotypical and anti-stereotypical, respectively. Following the principle of Counterfactual Data Augmentation (Zhao et al., 2018), we aim to artificially amplify or break associations between target words and their stereotypical attributes for obtaining our contrastive models. We use English Wikipedia as a base and filter the corpus for sentences $s_{(i,j)}$ containing exactly one target word $t \in T_i$ and one attribute word $a \in A_j$, where $i, j \in \{1, 2\}$. A sentence $s_{(i,j)}$ is categorized as stereotypical if $i = j$ and anti-stereotypical if $i \neq j$. For constructing a stereotypical dataset, we iterate through all sentences $s_{(i,j)}$ and retain those that are stereotypical. In cases where $s_{(i,j)}$ is anti-stereotypical, i.e. $i \neq j$, we replace the target term $t \in T_i$ with its corresponding paired term $t' \in T_j$. For creating an anti-stereotypical dataset, we keep all sentences $s_{(i,j)}$ with $i \neq j$ and substitute $t \in T_i$ with its paired target term $t' \in T_j$, if $i = j$. Note, that the resulting stereotypical and anti-stereotypical datasets are identical besides the swapped target terms.

4.1.3 Fine-tuning

We fine-tune BERT on the biased datasets using a masked language modeling (MLM) objective, creating stereotypical and anti-stereotypical models. To achieve higher levels of bias, we adjust the masking function to mask the target and attribute terms from our bias specification preferentially, i.e. with higher probability. We keep the average number of masked tokens constant by lowering the masking probability for all other tokens accordingly. We tested preferential masking probabilities between 0.15 and 0.9 and found that a value of 0.3 resulted in the best trade-off between perplexity and bias level. Additionally, we augmented the bi-

[1] we use the Huggingface BERT-base-uncased distribution.

[2] 20220301.en dump, Foundation

ased datasets with *neutral* examples not containing terms from the bias specification as this positively impacted the stability of subnetworks. To ensure the robustness of our findings, we conduct our experiments using four different random seeds. All training details can be found in appendix B.

4.2 Bias Evaluation

We measure gender bias using three well-established bias benchmarks, namely WEAT, StereoSet and CrowS-Pairs, all measuring intrinsic bias. The Word Embedding Association Test (WEAT) (Caliskan et al., 2017) measures the differential association of two sets of target words w.r.t. two sets of attribute words based on embedding similarity. We utilize the WEAT 8 test that compares male and female target terms to attribute terms related to art and science, respectively (see appendix D for the full specification). As many of these terms are also included in our bias specification in 4.1.1, WEAT 8 is a good instrument to precisely measure the gendered associations that we injected in our target and reference models and how these change when editing weights. We report the average effect size, where positive values indicate stereotypical bias and negative values indicate anti-stereotypical bias. We include a more detailed description of WEAT in appendix C.

We use the StereoSet intrasentence task (Nadeem et al., 2020) as a second bias measure, evaluating stereotypical bias at the sentence level across four target domains. StereoSet contains sentences describing a target group with a blank (e.g. *Girls tend to be more* `[BLANK]` *than boys*), which can be filled with a stereotypical (e.g. *soft*), an anti-stereotypical (e.g. *determined*) or a meaningless (e.g. *fish*) attribute. The *StereoSet stereotype (SS) score*, ranging from 0 to 100, is computed based on how often a model fills the blank with a stereotypical attribute over an anti-stereotypical attribute. A SS score >50 indicates stereotypical bias, while a score <50 indicates anti-stereotypical bias. We report scores for the gender-related portion of the dataset.

Our third measure, CrowS-Pairs (Nangia et al., 2020), comprises a crowd-sourced dataset of sentence pairs, with the first sentence being more stereotypical (e.g. *Women are always too sensitive about things*) than the second (e.g. *Men are always too sensitive about things*). The sentences differ only in the tokens related to the target group (e.g. *women/men*). Following Meade et al. (2022) we calculate the *CrowS-Pairs stereotype (CS) score* based on how often a model prefers the more stereotypical sentence. Scores range from 0 to 100, with CS scores >50 indicating a preference for the more stereotypical sentence and a CS score of <50 indicating the opposite. We report scores only for the gender-related portion of the dataset.

Both Stereoset and CrowS-Pairs have been criticized in the past for some of their properties (Blodgett et al., 2021; Pikuliak et al., 2023; Delobelle et al., 2022). We use these measures only complementary to WEAT and will later show that we observe similar effects across all three measures.

4.3 Subnetwork Identification

We discover subnetworks through weight pruning, inspired by the lottery ticket hypothesis (LTH) (Frankle and Carbin, 2018). The LTH posits that dense, randomly-initialized neural networks contain smaller subnetworks (*winning tickets*) that can be trained in isolation to achieve the accuracy of the full model. We use the approach of Chen et al. (2020) and apply iterative magnitude pruning (IMP) to extract subnetworks from the stereotypical and anti-stereotypical models. In IMP, we alternate between fine-tuning a model for i steps and subsequently pruning 10% of the weights with the lowest magnitude. After pruning, we reset the remaining weights to their initial values and repeat the steps until achieving a desired sparsity level. We accept a subnetwork as winning ticket if its performance after i fine-tuning steps is within 5% of the performance of the full model, fine-tuned for the same number of steps. Additional details on our application of IMP can be found in appendix E. Note that besides IMP, we also explored structured attention head pruning (Prasanna et al., 2020) but found no differences between stereotypical and anti-stereotypical subnetwork masks. We suspect that this occurred due to attention heads being too coarse-grained to capture the subtle differences in bias we injected.

4.4 Uninformed Editing

To explore the importance of the localization step, we also deploy our editing strategies to subsets of weights that are not informed by strategic localization. For *uninformed intrapolation* and *uninformed extrapolation*, we randomly sample from all weights, excluding those pruned in both the target and reference subnetworks, as intra- or extrapolation would not affect these weights. For

	full	10%	20%	30%	40%
WEAT	0.93	0.88	0.98	0.89	0.91
StereoSet	61.03	60.27	60.09	61.11	60.95
CrowS-Pairs	57.92	59.45	59.83	57.25	58.88

(a) stereotypical subnetworks

	full	10%	20%	30%	40%
WEAT	-0.75	-0.75	-0.56	-0.63	-0.44
StereoSet	58.86	58.37	58.03	59.52	59.30
CrowS-Pairs	52.77	53.15	52.86	52.01	52.39

(b) anti-stereotypical subnetworks

Table 1: **Bias of subnetworks at different sparsities**. We report the mean across all random seeds with higher scores indicating higher stereotypical bias.

uninformed pruning, we randomly sample from the weights that are present in the target subnetwork, excluding those that already are pruned. To ensure a fair comparison, we select the same number of weights as those identified by the localization strategies in section 3.1.

5 Results

We fine-tune BERT according to section 4.1 and create models with stereotypical and anti-stereotypical biases. Table 1 ("full") shows that the stereotypical models exhibit higher levels of measurable stereotypical bias than the anti-stereotypical models. Thus, as intended, we have successfully steered the BERT model in two extreme directions, which will serve as a basis for our experiments on contrastive editing. We observe a trend for both types of models to shift towards the stereotypical regime, due to the fact that the bias benchmarks test for a broader range of associations than those we artificially controlled. This effect is less pronounced for WEAT, as WEAT specifically tests for many of our injected associations.

5.1 Subnetwork Analysis

We discover subnetworks that are winning tickets (cf. section 4.3) for both stereotypical and anti-stereotypical models at different sparsities up to 40%. The discovered subnetworks are stable across runs with different random seeds, as indicated by a high Jaccard similaritiy of the pruning masks (> 0.98). This suggests that the findings are robust and not heavily influenced by factors such as specific data splits. Table 1 illustrates that the discovered subnetworks largely maintain the bias of their parent networks, highlighting their suitability for our contrastive approach.

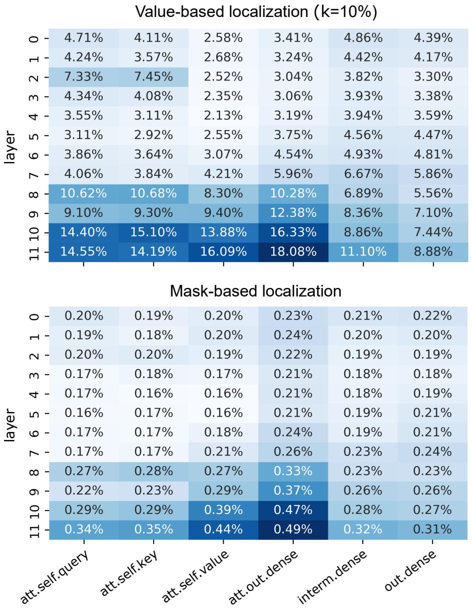


Figure 3: **Bias localization.** We illustrate the percentage of weights per component that have been selected for editing. Notably, both localization strategies focus on the same layers and components. We show the results for subnetworks at sparsity 40% and one random seed.

Next, we compare subnetworks with stereotypical bias to subnetworks with anti-stereotypical bias. At all sparsity levels, we find that the percentage of weights where the pruning masks differ remains below 0.5%, indicating a high degree of similarity. This is expected, as both types of subnetworks specialize on the same task and are fine-tuned on similar datasets, differing only in the injected stereotypical and anti-stereotypical associations, respectively. To investigate where these associations manifest in the parameter space, we apply both mask-based and value-based localization. Although the localization strategies are designed not to select the same subsets of weights, they consistently target similar areas within the model, particularly focusing on the last layers and predominantly the attention output dense layer (see figure 3). This observation aligns with previous work (Ma et al., 2023; Chintam et al., 2023), which found that attention heads most influential for gender bias are located in higher layers, suggesting that bias is encoded in specific subcomponents of transformers.

5.2 Effect on Gender Bias

We investigate the effectiveness of the local contrastive editing strategies by considering two settings: using the stereotypical subnetworks as the target with the anti-stereotypical subnetworks as the reference, and vice versa. Figure 4 illustrates results for selected parameter settings, demonstrating the flexibility of our strategies.

We find that nearly all editing strategies, when combined with either mask- or value-based localization, effectively modify gender bias as intended. Mask-based editing achieves this efficiently with a small subset size of less than 0.5%. By varying the weighting factor α, we can flexibly control the bias of the target model and, according to WEAT, even completely remove bias at $\alpha = 2$. In contrast, uninformed edits result in minimal or no changes to the bias scores, highlighting the critical role of the localization step in local contrastive editing. We summarize that (i) both mask-based and value-based localization can identify subsets of weights driving stereotypical gender bias, and that (ii) gender bias can be controlled through contrastive editing strategies on these subsets.

5.3 Effect on Performance

We further examine how local contrastive editing affects a model's ability to model language. To assess this, we use perplexity as a primary measure and additionally compute the *language modeling (LM) score* as introduced by Nadeem et al. (2020). Similar to the SS score (cf. section 4.2), the LM score is computed on the StereoSet dataset and measures how frequently a model prefers a sentence with a meaningful association (e.g. *Girls tend to be more **determined** than boys*) over a nonsensical one (e.g. *Girls tend to be more **fish** than boys*). The LM score ranges from 0 to 100, where 100 represents an ideal model that always favours the semantically meaningful association.

As shown in figure 4, nearly all editing strategies lead to only minor increases in perplexity. The notable exception is value-based pruning, which causes a significant increase of 9.529 points in perplexity, while extrapolation also leads to a significant but much smaller rise. This substantial degradation in model performance may explain the counterintuitive effect of value-based pruning on gender bias in figure 4, as the models' overall functionality is severely impaired. Consistent with these findings, the LM score shows no significant decline, except for value-based pruning, which leads to a significant drop of 2.723 points at a sparsity level of 30%. Complete LM score results and results for other sparsity levels can be found in appendix F.2.

Next, we fine-tune the edited models on the MNLI and STS-B tasks from the GLUE benchmark (Wang et al., 2018), to evaluate their performance on downstream tasks. As shown in table 2, the model subjected to value-based pruning again exhibits the most significant performance drop compared to the base model [3]. Models edited using other strategies experience a maximum performance loss of only $0.71\%/0.59\%$ for MNLI and $1.07\%/1.06\%$ for STS-B.

Model	MNLI-m/mm	STS-B
Base	**84.4/84.8**	**89.0/88.9**
IP (α=0.5)	83.9/84.2	88.6/88.0
IP (α=1)	83.9/84.4	88.5/88.0
EP (α=2)	83.9/84.3	88.5/88.0
EP (α=-2)	83.8/84.4	88.5/88.1
PR	83.9/84.3	88.4/87.9
SW	83.8/84.4	88.5/88.0

(a) mask-based loc.

Model	MNLI-m/mm	STS-B
Base	**84.4/84.8**	**89.0/88.9**
IP (α=0.5)	83.9/84.3	88.6/88.0
IP (α=1)	83.8/84.3	88.5/88.0
EP (α=2)	84.0/84.4	88.5/87.9
EP (α=-2)	83.8/84.5	88.6/88.1
PR	83.5/83.8	87.0/86.6

(b) value-based loc. (k=10%)

Table 2: **Performance on downstream tasks**. We fine-tune the edited models on the MNLI and STS-B tasks from the GLUE benchmark and show the results for the stereotypical target model at 30% sparsity using a single random seed. For MNLI, we report both matched and mismatched accuracy while for STS-B, we present Pearson and Spearman correlation. Results for other sparsities and target models can be found in the appendix F.3.

In summary, we demonstrate that, with exception of value-based pruning, local contrastive editing largely preserves a model's language modeling ability. Furthermore, the edited models can still be effectively used in downstream tasks with only minor performance loss.

5.4 Ablation Study

We explore the impact of the number of selected weights k, the weighting factor α, and the different localization strategies on inter- and extrapolation.

Number of Selected Weights Figure 5 shows the sensitivity of value-based interpolation ($\alpha = 0.5$) to the number of selected weights k. Increasing k initially leads to stronger effects on gender bias up to a threshold of 20–40%. Beyond this range, editing further weights seems to have no effect, indicating that there is a critical subset primarily encoding the bias.

Weighting Factor Figure 6 shows the behavior of linear weight inter- and extrapolation for all localization strategies. For WEAT, we can achieve a smooth, monotonous change in gender bias by gradually increasing α. StereoSet and CrowS-Pairs

[3] BERT-base-uncased

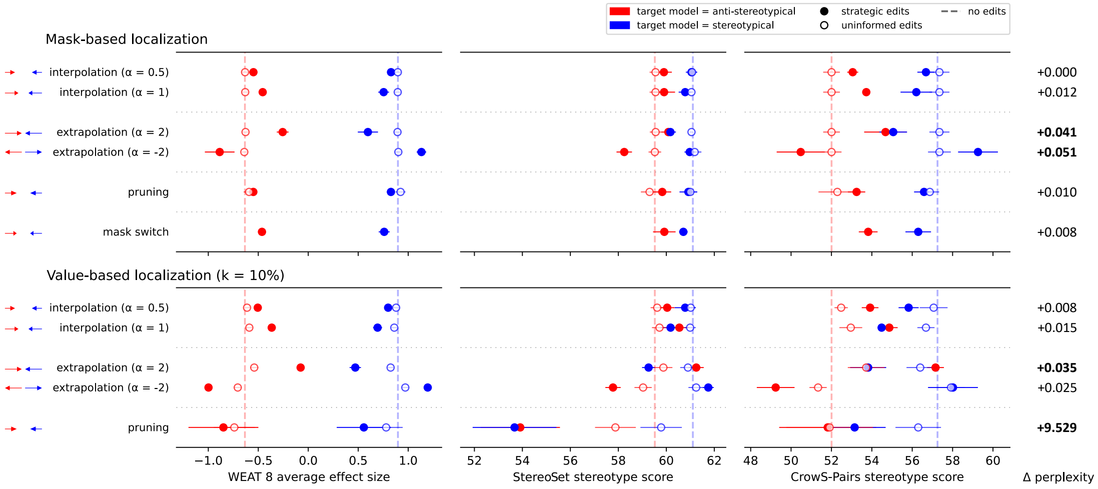


Figure 4: **Local contrastive editing of gender bias.** We illustrate the effects of our local editing strategies on gender bias for settings in which (i) the target model is anti-stereotypical and the reference is stereotypical (red) and (ii) the target model is stereotypical and the reference is anti-stereotypical (blue). The colored arrows on the left indicate the intuition of each strategy. We show the results for subnetworks at sparsity 30% and report the mean bias over four random seeds, with error bars indicating one standard deviation in each direction. On the right, we display the mean perplexity change across both settings and all random seeds where bold indicates a significant increase. Our local editing strategies can successfully steer stereotypical bias with both localization methods, while uninformed edits have much lower or no effect at all. Results for other sparsities can be found in appendix F.1.

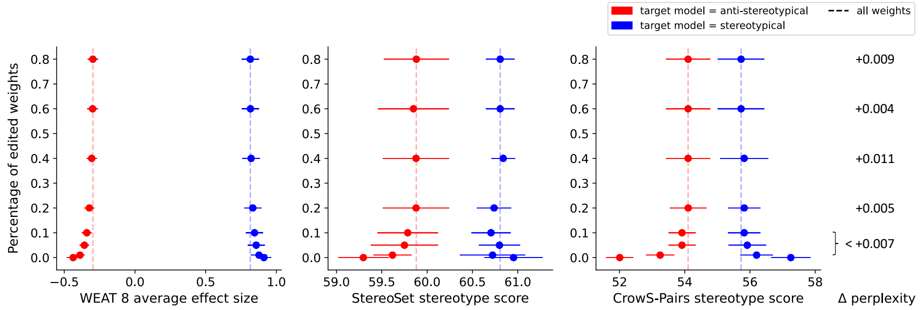
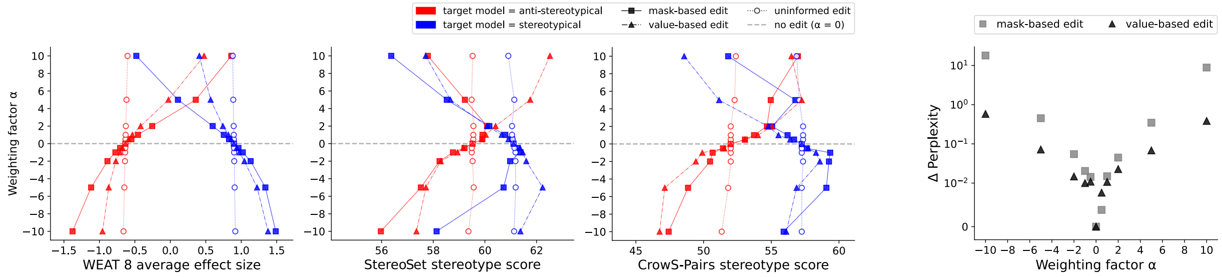


Figure 5: **Sensitivity to the number of weights edited.** We explore the influence of the number of top-k weights that are used for value-based interpolation with $\alpha = 0.5$. We show the results for subnetworks at sparsity 30% and report the mean bias and standard deviation across four random seeds. We report the average change in perplexity across both target models and all random seeds, observing no significant increase for any choice of k. Results for other sparsities can be found in appendix F.4.



(a) Gender bias

(b) Perplexity

Figure 6: **Sensitivity to the weighting factor.** We investigate the effect of different weighting factors α on gender bias (a) and perplexity (b). For value-based and uninformed edits we choose the same number of weights that were selected by mask-based localization. We find that weighting factors with higher magnitudes lead to greater effects on bias, correlating with an increase in perplexity. We display the results for a sparsity level of 30% and report the mean across four random seeds. For perplexity, we average the results for both target models. Results for other sparsities can be found in appendix F.5.

measure a similar trend, but the effects become inconsistent for higher absolute weighting factors ($|\alpha| \geq 5$), likely due to a decline in language modeling performance, as evidenced by increasing perplexity (see figures 6b, 17) and decreasing LM scores (see figure 18). Overall, we find that varying α allows flexible control and calibration of bias levels within a target model, that can be tailored to the characteristics of the reference model (e.g. reducing bias when the reference model itself exhibits bias), albeit within certain limits.

Localization Strategies We choose the same number of weights to be edited for all localization strategies, allowing their direct comparison in figure 6. We observe that the localization strategy that leads to stronger steering effects as measured by WEAT (specifically mask-based localization at 30% sparsity) also leads to a stronger decline in language modeling ability when $|\alpha|$ increases. In line with the result in figure 4, we further find that uninformed editing does not change the bias level significantly, not even for high magnitudes of α. This suggests that certain subsets of weights encode gender bias more prominently, and that their localization can be crucial for bias modification.

5.5 Wider Applicability

So far, our experiments have been conducted in a controlled environment where the target and reference models were fine-tuned on parallel datasets. To test the wider applicability of our approach, we fine-tune a *neutral* model on a subset of Wikipedia that is independent of the biased datasets described in section 4.1. The training details can be found in appendix B. In line with the other experiments, we extract neutral subnetworks at different sparsities with four random seeds. We use those as target networks and edit them w.r.t both stereotypical and anti-stereotypical reference models. Figure 7 illustrates the results at sparsity 40%. By using the stereotypical reference model we can successfully modify the bias of the neutral model in line with our intuition. For instance, extrapolation with $\alpha = -2$, successfully removes stereotypical bias, as measured by WEAT. This is not trivial, as here the reference and target models are fine-tuned on datasets that do not overlap, which implies that the weights selected by the localization strategies may encode differences in the datasets beyond just stereotypical bias. Using an anti-stereotypical reference model produces mixed results, aligning with expectations in most but not all scenarios, requiring further investigation.

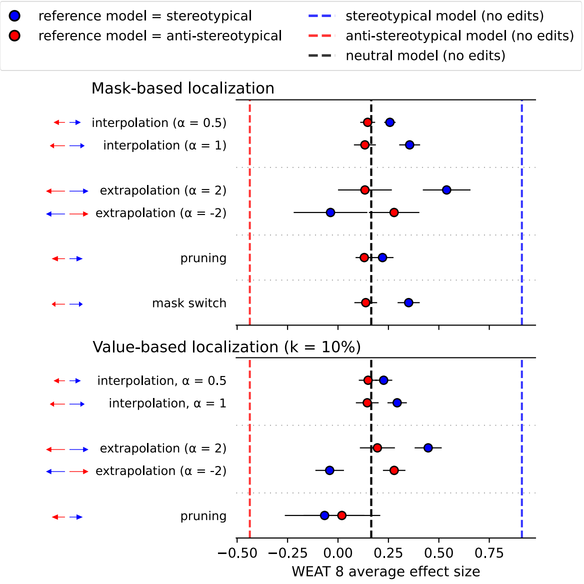


Figure 7: **Application to a neutral model.** We apply local contrastive editing to a *neutral* target model. Using a stereotypical reference model (blue) can effectively steer the neutral model's bias. Using an anti-stereotypical reference model (red) produces inconclusive results. We show the results for a sparsity of 40% and report the mean and standard deviation across four random seeds. Results for other sparsities can be found in appendix F.6

6 Conclusion

Our research shows that stereotypical gender bias is primarily encoded in specific subsets of weights within LMs. We propose various local contrastive editing strategies and demonstrate that they can effectively identify and modify these subsets to flexibly control and mitigate gender bias. This work enhances our understanding of where stereotypical biases manifest in the parameter space of LMs and opens up new avenues for developing parameter-efficient strategies for model editing in a contrastive manner. Local contrastive editing is not limited to gender bias, and future research could explore its application to other tasks and domains.

7 Limitations

Naturally, our work comes with limitations. We conduct experiments using a single model architecture and a single bias type only. We restrict our study to this model architecture because of computational constraints and environmental considerations, particularly because iterative magnitude pruning requires substantial computational resources. However, we anticipate that our findings generalize broadly, as related work on weight averaging has been shown to generalize to other model architectures as well (Wortsman et al., 2022; Yadav et al., 2024; Ilharco et al., 2022b). Moreover, findings from other studies on bias mitigation suggest generalizability to other types of bias that share similar specifications (Hauzenberger et al., 2023; Guo et al., 2022).

8 Ethical Considerations

While our work ultimately targets the development of strategies for reducing bias in language models, it is important to acknowledge the potential dual use of these techniques. The same strategies designed to reduce bias, can also be used to perpetuate and amplify biases in language models. Moreover, our intentional design of language models that initially exhibit high levels of bias raises concerns about their potential deployment in various applications. Despite our intention to mitigate bias, the existence of such models risks normalizing and perpetuating stereotypical gender biases in society.

Additionally, our study focuses on bias specifically between male and female groups, thus operating on a binary specification of gender bias, only. We recognize that gender is a spectrum, encompassing identities beyond strictly male and female categories. Our rationale for this approach is to evaluate the effectiveness of our strategies using an existing and well-established specification that facilitates measurement of their effects.

Acknowledgments

We thank Ekaterina Shutova for her valuable input and feedback for this project. The authors acknowledge support by the state of Baden-Württemberg through bwHPC and the German Research Foundation (DFG) through grant INST 35/1597-1 FUGG. The work of Anne Lauscher is funded under the Excellence Strategy of the German Federal Government and the Federal States. The work of Rochelle Choenni is supported by a Google PhD Fellowship.

References

Soumya Barikeri, Anne Lauscher, Ivan Vulić, and Goran Glavaš. 2021. RedditBias: A real-world resource for bias evaluation and debiasing of conversational language models. 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 1941–1955, Online. Association for Computational Linguistics.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

Su Lin Blodgett, Gilsinia Lopez, Alexandra Olteanu, Robert Sim, and Hanna Wallach. 2021. Stereotyping Norwegian salmon: An inventory of pitfalls in fairness benchmark datasets. 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 1004–1015, Online. Association for Computational Linguistics.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29.

Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Tianlong Chen, Jonathan Frankle, Shiyu Chang, Sijia Liu, Yang Zhang, Zhangyang Wang, and Michael Carbin. 2020. The lottery ticket hypothesis for pretrained bert networks. *Advances in neural information processing systems*, 33:15834–15846.

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.

Rochelle Choenni, Dan Garrette, and Ekaterina Shutova. 2023a. Cross-lingual transfer with language-specific subnetworks for low-resource dependency parsing. *Computational Linguistics*, pages 613–641.

Rochelle Choenni, Ekaterina Shutova, and Dan Garrette. 2023b. Examining modularity in multilingual lms via language-specialized subnetworks. *arXiv preprint arXiv:2311.08273*.

Rochelle Choenni, Ekaterina Shutova, and Robert van Rooij. 2021. Stepmothers are mean and academics are pretentious: What do pretrained language models learn about you? In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1477–1491.

Pieter Delobelle, Ewoenam Tokpo, Toon Calders, and Bettina Berendt. 2022. Measuring fairness with biased rulers: A comparative study on bias metrics for pre-trained language models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1693–1706, Seattle, United States. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Xiangjue Dong, Yibo Wang, Philip S Yu, and James Caverlee. 2024. Disclosure and mitigation of gender bias in llms. *arXiv preprint arXiv:2402.11190*.

Wikimedia Foundation. Wikimedia downloads.

Jonathan Frankle and Michael Carbin. 2018. The Lottery Ticket Hypothesis: Finding Sparse, Trainable Neural Networks. In *International Conference on Learning Representations*.

Hila Gonen and Yoav Goldberg. 2019. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 609–614.

Anthony G Greenwald, Debbie E McGhee, and Jordan LK Schwartz. 1998. Measuring individual differences in implicit cognition: the implicit association test. *Journal of personality and social psychology*, 74(6):1464.

Almog Gueta, Elad Venezian, Colin Raffel, Noam Slonim, Yoav Katz, and Leshem Choshen. 2023. Knowledge is a region in weight space for fine-tuned language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1350–1370, Singapore. Association for Computational Linguistics.

Yue Guo, Yi Yang, and Ahmed Abbasi. 2022. Autodebias: Debiasing masked language models with automated biased prompts. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1012–1023, Dublin, Ireland. Association for Computational Linguistics.

Lukas Hauzenberger, Shahed Masoudian, Deepak Kumar, Markus Schedl, and Navid Rekabsaz. 2023. Modular and on-demand bias mitigation with attribute-removal subnetworks. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6192–6214, Toronto, Canada. Association for Computational Linguistics.

Amr Hendy, Mohamed Abdelghaffar, Mohamed Afify, and Ahmed Y Tawfik. 2022. Domain Specific Subnetwork for Multi-Domain Neural Machine Translation. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing*, pages 351–356.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. 2022a. Editing models with task arithmetic. In *The Eleventh International Conference on Learning Representations*.

Gabriel Ilharco, Mitchell Wortsman, Samir Yitzhak Gadre, Shuran Song, Hannaneh Hajishirzi, Simon Kornblith, Ali Farhadi, and Ludwig Schmidt. 2022b. Patching open-vocabulary models by interpolating weights. *Advances in Neural Information Processing Systems*, 35:29262–29277.

Przemyslaw Joniak and Akiko Aizawa. 2022. Gender biases and where to find them: Exploring gender bias in pre-trained transformer-based language models using movement pruning. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 67–73, Seattle, Washington. Association for Computational Linguistics.

Anne Lauscher, Goran Glavaš, Simone Paolo Ponzetto, and Ivan Vulić. 2020. A general framework for implicit and explicit debiasing of distributional word vector spaces. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8131–8138.

Anne Lauscher, Tobias Lueken, and Goran Glavaš. 2021. Sustainable modular debiasing of language models. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4782–4797, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Susan Leavy, Gerardine Meaney, Karen Wade, and Derek Greene. 2020. Mitigating gender bias in machine learning data sets. In *Bias and Social Aspects in Search and Recommendation: First International Workshop, BIAS 2020, Lisbon, Portugal, April 14, Proceedings 1*, pages 12–26. Springer.

Weicheng Ma, Henry Scheible, Brian Wang, Goutham Veeramachaneni, Pratim Chowdhary, Alan Sun, Andrew Koulogeorge, Lili Wang, Diyi Yang, and Soroush Vosoughi. 2023. Deciphering stereotypes in pre-trained language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11328–11345, Singapore. Association for Computational Linguistics.

Nicholas Meade, Elinor Poole-Dayan, and Siva Reddy. 2022. An empirical survey of the effectiveness of debiasing techniques for pre-trained language models. In *Proceedings of the 60th Annual Meeting of the*

Association for Computational Linguistics (Volume 1: Long Papers), pages 1878–1898, Dublin, Ireland. Association for Computational Linguistics.

Johannes Mario Meissner, Saku Sugawara, and Akiko Aizawa. 2022. Debiasing masks: A new framework for shortcut mitigation in NLU. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7607–7613, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2020. Stereoset: Measuring stereotypical bias in pretrained language models. *arXiv preprint arXiv:2004.09456*.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel Bowman. 2020. Crows-pairs: A challenge dataset for measuring social biases in masked language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1953–1967.

Farhad Nooralahzadeh and Rico Sennrich. 2023. Improving the cross-lingual generalisation in visual question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 13419–13427.

Matúš Pikuliak, Ivana Beňová, and Viktor Bachratý. 2023. In-depth look at word filling societal bias measures. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3648–3665, Dubrovnik, Croatia. Association for Computational Linguistics.

Sai Prasanna, Anna Rogers, and Anna Rumshisky. 2020. When BERT Plays the Lottery, All Tickets Are Winning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3208–3229, Online. Association for Computational Linguistics.

Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating gender bias in natural language processing: Literature review. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1630–1640.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. 2020. Investigating gender bias in language models using causal mediation analysis. *Advances in neural information processing systems*, 33:12388–12401.

Ivan Vulić, Simon Baker, Edoardo Maria Ponti, Ulla Petti, Ira Leviant, Kelly Wing, Olga Majewska, Eden Bar, Matt Malone, Thierry Poibeau, et al. 2020. Multi-simlex: A large-scale evaluation of multilingual and crosslingual lexical semantic similarity. *Computational Linguistics*, 46(4):847–897.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Zirui Wang, Zachary C Lipton, and Yulia Tsvetkov. 2020. On negative interference in multilingual models: Findings and a meta-learning treatment. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4438–4450.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Mitchell Wortsman, Gabriel Ilharco, Samir Ya Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, et al. 2022. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *International conference on machine learning*, pages 23965–23998. PMLR.

Prateek Yadav, Derek Tam, Leshem Choshen, Colin A Raffel, and Mohit Bansal. 2024. Ties-merging: Resolving interference when merging models. *Advances in Neural Information Processing Systems*, 36.

Yunzhi Yao, Peng Wang, Bozhong Tian, Siyuan Cheng, Zhoubo Li, Shumin Deng, Huajun Chen, and Ningyu Zhang. 2023. Editing large language models: Problems, methods, and opportunities. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10222–10240, Singapore. Association for Computational Linguistics.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. 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 15–20, New Orleans, Louisiana. Association for Computational Linguistics.

A Gender Bias Specification

Target Word Pairs

$T_1 \times T_2$: *(woman, man), (women, men), (girl, boy), (mother, father), (daughter, son), (wife, husband), (niece, nephew), (mom, dad), (bride, groom), (lady, gentleman), (madam, sir), (hostess, host), (female, male), (wife, husband), (aunt, uncle), (sister, brother), (she, he)*

Attribute Words

A_1: *nurse, secretary, housekeeper, cook, sewing, cleaning, waitress, maid, child-care, receptionist, clerk, home-care, poetry, art, dance, literature, novel, symphony, drama, sculpture, shakespeare*

A_2: *surgeon, executive, manager, officer, engineering, programming, lawyer, engineer, finance, administrator, physician, science, math, geometry, technology, equation, computation, physics, chemistry, einstein*

B Training Details

For fine-tuning and pruning we used 4 NVIDIA A100-80GB GPUs. One pruning and fine-tuning iteration took 32 GPU hours, amounting to 160 GPU hours to extract subnetworks up to sparsity 40% from a single model. As we repeated this for three types of models and four random seeds, we amount in a total of 1920 GPU hours.

	Biased	**Neutral**
# Biased examples	164,524	0
# Neutral examples	164,524	329,048
Task	MLM	MLM
Masking prob.	0.3 / 0.148	0.15
# Epochs	3	3
# Iterations/epoch	4627	4627
Batch size	64	64
Learning rate	1×10^{-5}	1×10^{-5}
Eval size	0.1	0.1
Eval Measure	Perplexity	Perplexity
# Random seeds	4	4

Table 3: Details of fine-tuning biased (stereotypical or anti-stereotypical) and neutral models. Optimization is performed with AdamW with $\epsilon = 1 \times 10^{-8}$ and the learning rate decays linearly to zero. We use standard implementations and hyperparameter settings (Wolf et al., 2020).

C Word Embedding Association Test (WEAT)

Caliskan et al. (2017) introduce WEAT by extending the Implicit Association Test (Greenwald et al., 1998), a test used to measure human biases, to word embeddings. The test measures the differential association of two sets of target words X, Y (e.g. *female* and *male* terms) w.r.t. two sets of attribute words A, B (e.g. *art* and *science* terms) based on their cosine similarity in the embedding space:

$$s(X,Y,A,B) = \sum_{x \in X} s(x,A,B) - \sum_{y \in Y} s(y,A,B)\,,$$

where

$$s(w,A,B) = \frac{1}{|A|}\sum_{a \in A} cos(w,a) - \frac{1}{|B|}\sum_{b \in B} cos(w,b)\,.$$

The significance of the test is computed with a permutation test, where $\{(X_i, Y_i)\}_i$ are all equally sized partitions of $X \cup Y$ into two sets:

$$Pr_i((s(X_i,Y_i,A,B) > s(X,Y,A,B))$$

Here, we report the effect size as a measure of separation between the association distributions:

$$\frac{\text{mean}_{x \in X} s(x,A,B) - \text{mean}_{y \in Y} s(y,A,B)}{\text{std-dev}_{w \in X \cup Y} s(w,A,B)}$$

Following Vulić et al. (2020), we extract embeddings for all target and attribute words by feeding them through BERT, prepended with the start of sequence token and appended with the separator token (e.g. `[CLS] woman [SEP]`). We then extract embeddings from each hidden layer and compute WEAT separately for each layer. Finally, we report the average effect size of the test across all layers.

D WEAT 8 Specification

Target Words

X: *science, technology, physics, chemistry, Einstein, NASA, experiment, astronomy*

Y: *poetry, art, Shakespeare, dance, literature, novel, symphony, drama*

Attribute Words

A: *brother, father, uncle, grandfather, son, he, his, him*

B: *sister, mother, aunt, grandmother, daughter, she, hers, her*

E Iterative Magnitude Pruning

We apply iterative magnitude pruning according to the following procedure:

1. Fine-tune a pre-trained network $f(\cdot, \theta_0)$ for i steps
2. Globally prune $p\%$ of the weights with the lowest magnitude, resulting in a subnetwork $f(\cdot, m \odot \theta_i)$ with pruning mask m
3. Reset the remaining weights to their initial values θ_0
4. Repeat the previous steps on $f(\cdot, m \odot \theta_0)$ until the desired sparsity level is reached

We set the pruning rate per iteration to $p\% = 10\%$. Consistent with Chen et al. (2020), we only prune weights (and e.g. not biases) and exclude the embedding layer and the task-specific layer from the pruning process. In each fine-tuning iteration, we use the number of steps and parameter settings detailed in appendix B.

F Additional Results

F.1 Effect of Local Contrastive Editing on Gender Bias

We show the effects of local contrastive editing on gender bias for additional sparsity levels in figures 8, 9 and 10. We observe that as the sparsity level increases, the impact of mask-based editing becomes more pronounced, whereas the influence of value-based editing on bias declines.

F.2 Language Modeling Ability

We present the language modeling ability (as measured by perplexity and LM score) of the discovered subnetworks at varying sparsity levels in table 4. Both, perplexity and LM score are comparable between stereotypical and anti-stereotypical subnetworks and remain similar across different sparsity levels.

Table 5 presents the change in language modeling ability after local contrastive editing. We present the average changes in perplexity and LM score across both target models and four random seeds at different sparsity levels. We apply a one-sided Wilcoxon signed-rank test with Bonferroni correction to assess whether the observed changes (an increase in perplexity or a decrease in LM scores) are significant. Notably, across all sparsity levels, perplexity remains consistently low, with significant increases occurring occasionally after extrapolation with large absolute weighting factors, and consistently across all sparsities for value-based pruning. A similar trend is observed for LM score, with significant drops occurring only in the case of value-based pruning.

	full	10%	20%	30%	40%
Perplexity	5.57	5.56	5.59	5.69	5.88
LM score	85.58	85.51	85.80	85.81	85.53

(a) stereotypical subnetworks

	full	10%	20%	30%	40%
Perplexity	5.57	5.58	5.60	5.68	5.90
LM score	85.88	85.63	85.89	85.78	85.51

(b) anti-stereotypical subnetworks

Table 4: **Language modeling ability of subnetworks at different sparsities**. We report the mean across all random seeds, where *lower* perplexity and *higher* LM scores indicate better performance in terms of language modeling.

F.3 Downstream Tasks

Tables 6 and 7 present the results of fine-tuning the edited models on the MNLI and STS-B tasks from the GLUE benchmark for different sparsity levels. The edited models exhibit only a slight decrease in performance compared to the base model, with the lowest performing models achieving $82.9/83.4$ $(-1.8\%/-1.2\%)$ on MNLI and $86.6/86.24$ $(-2.7\%/-3.0\%)$ on STS-B. We note that at each sparsity level, the models with the greatest performance drop were those subjected to value-based pruning prior to fine-tuning on the downstream task.

F.4 Sensitivity to the Number of Weights Edited

Figures 11, 12 and 13 present the results of experiments exploring various numbers of weights selected for interpolation for additional sparsity levels, using a fixed weighting factor of $\alpha = 0.5$. We observe similar trends across all bias measures and sparsity levels, indicating that editing $20\% - 40\%$ of the weights has an equivalent effect on bias as editing all the weights.

F.5 Sensitivity to the Weighting Factor

We display results exploring different weighting factors for interpolation and extrapolation across

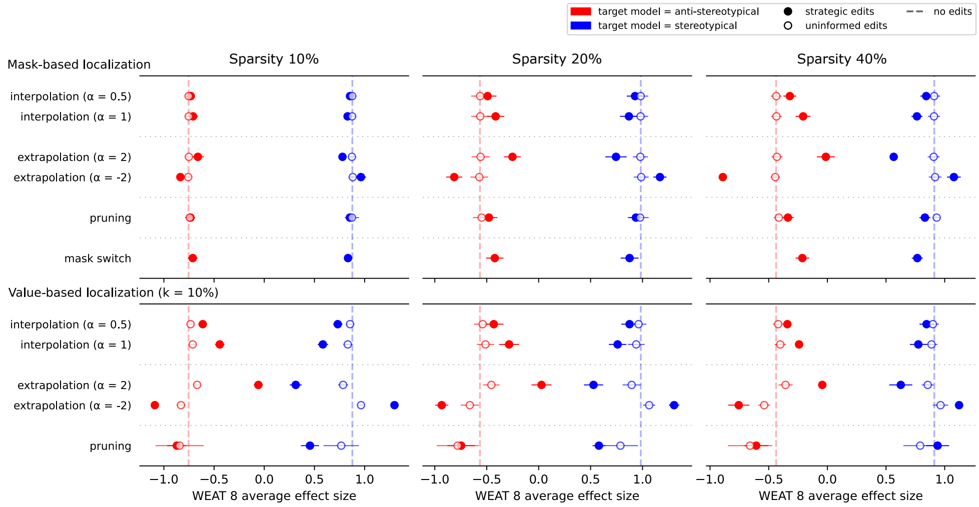


Figure 8: **WEAT average effect size after local contrastive editing.** We report the mean bias at different sparsity levels across four random seeds with error bars indicating one standard deviation in each direction.

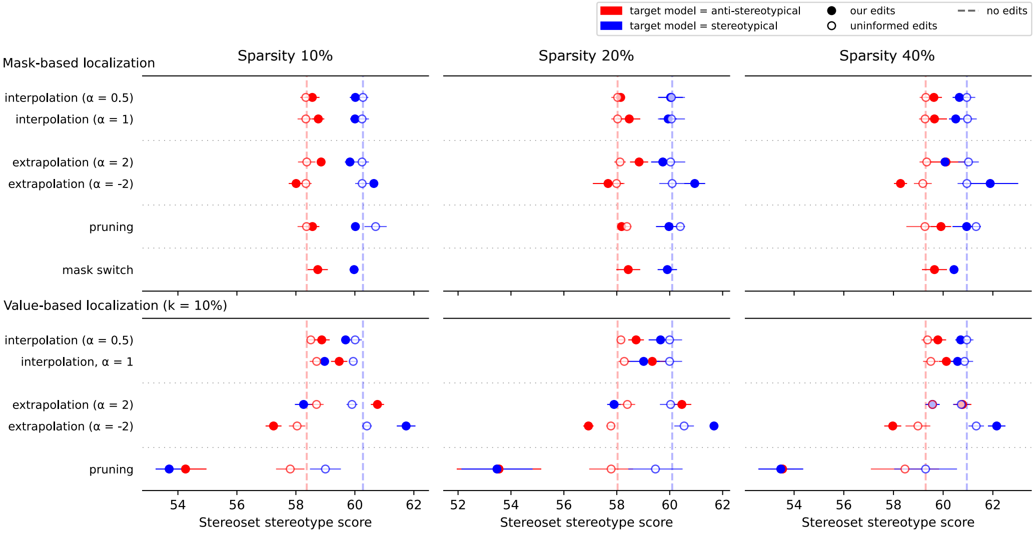


Figure 9: **StereoSet stereotype scores after local contrastive editing.** We illustrate the mean bias at different sparsity levels across four random seeds with error bars indicating one standard deviation in each direction.

different sparsity levels. The effect on gender bias is illustrated in figures 14, 15 and 16, while the effect on language modeling performance is shown in figures 17 and 18. We observe that across all sparsity levels and bias measures, the effect of mask-based and value-based editing on bias increases with higher absolute weighting factors (up to a certain threshold), whereas uninformed editing does not lead to any or only minor changes in bias. At the same time, perplexity and LM score indicate increasingly worse language modeling performance for higher absolute weighting factors.

F.6 Application to a Neutral Model

We show the effect of local contrastive editing on gender bias for a neutral target model at different sparsity levels in figures 19, 20 and 21. We record the change in language modeling ability in table 8.

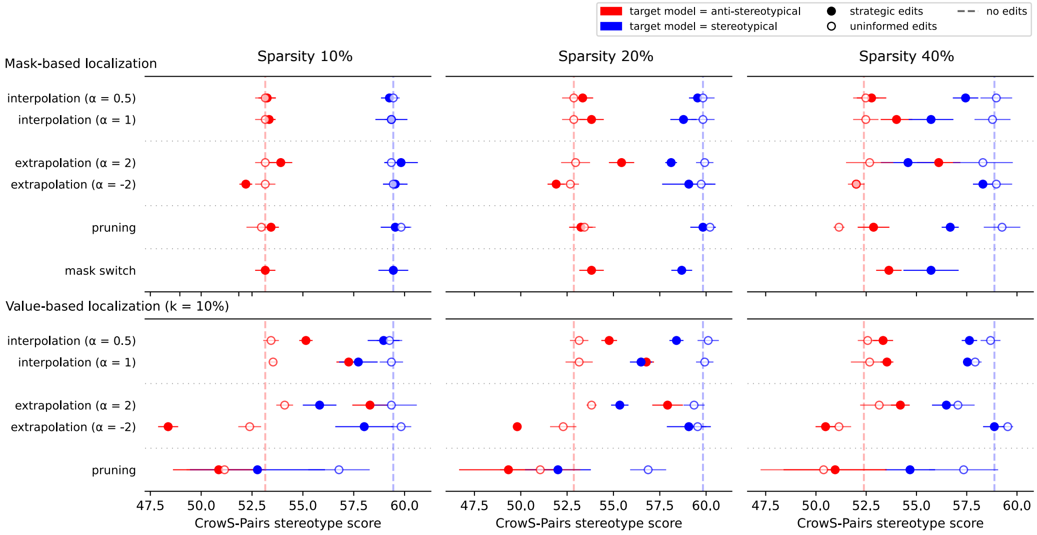


Figure 10: **CrowS-Pairs stereotype scores after local contrastive editing.** We show the mean bias at different sparsity levels across four random seeds with error bars indicating one standard deviation in each direction.

	Δ perplexity ↓				Δ LM score ↑			
	10%	20%	30%	40%	10%	20%	30%	40%
Mask-based localization								
IP ($\alpha = 0.5$)	-0.002	+0.001	+0.000	+0.007	+0.012	-0.010	+0.012	-0.009
IP ($\alpha = 1$)	-0.003	+0.001	+0.012	**+0.017**	-0.080	-0.045	-0.080	-0.038
EP ($\alpha = 2$)	+0.006	+0.014	**+0.041**	**+0.072**	+0.005	-0.016	+0.005	+0.015
EP ($\alpha = -2$)	+0.002	+0.016	**+0.051**	**+0.137**	-0.023	-0.050	-0.023	-0.099
PR	-0.003	-0.001	+0.010	+0.014	-0.005	-0.023	-0.005	-0.017
SW	+0.000	+0.010	+0.008	+0.019	-0.055	-0.061	-0.055	-0.023
Value-based localization ($k = 10\%$)								
IP ($\alpha = 0.5$)	-0.003	+0.006	+0.008	+0.002	-0.004	+0.013	-0.004	-0.010
IP ($\alpha = 1$)	+0.010	+0.013	+0.015	**+0.016**	-0.076	-0.017	-0.076	+0.005
EP ($\alpha = 2$)	**+0.051**	**+0.032**	**+0.035**	**+0.032**	-0.147	-0.048	-0.147	+0.020
EP ($\alpha = -2$)	+0.031	+0.020	+0.025	+0.017	+0.005	-0.233	+0.005	+0.059
PR	**+9.225**	**+10.18**	**+9.529**	**+11.722**	**-2.724**	**-3.500**	**-2.723**	**-0.965**

Table 5: **Change in language modeling ability.** We show the mean change in perplexity and LM score after local contrastive editing across both target models (stereotypical and anti-stereotypical) and four random seeds at different sparsity levels. We print significant differences bold.

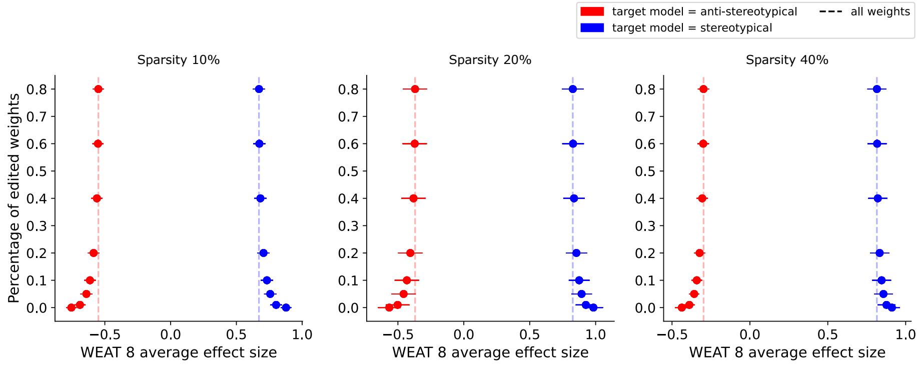


Figure 11: **WEAT average effect size for different numbers of weights edited.** We report the mean bias at different sparsity levels across four random seeds with error bars indicating one standard deviation in each direction.

	MNLI-m/mm ↑			STS-B ↑		
	10%	20%	40%	10%	20%	40%
Base		**84.4/84.8**			**89.0/88.9**	
Mask-based loc.						
IP ($\alpha = 0.5$)	84.6/84.5	84.4/84.6	84.1/84.0	88.7/88.3	88.7/88.4	88.2/87.6
IP ($\alpha = 1$)	84.5/84.5	84.3/84.5	84.0/84.1	88.7/88.3	88.8/88.4	88.2/87.6
EP ($\alpha = 2$)	84.6/84.6	84.3/84.5	84.1/84.0	88.7/88.3	88.8/88.3	88.2/87.6
EP ($\alpha = -2$)	84.6/84.6	84.3/84.5	83.9/84.0	89.0/88.7	89.0/88.5	88.1/87.6
PR	84.6/84.6	84.0/84.4	84.0/84.1	88.8/88.4	88.8/88.4	88.3/87.7
SW	84.7/84.6	84.2/84.5	84.0/84.2	88.8/88.4	88.9/88.4	88.2/87.6
Value-based loc. (k=10%)						
IP ($\alpha = 0.5$)	84.5/84.5	84.4/84.4	84.0/84.0	88.9/88.5	88.8/88.4	88.2/87.7
IP ($\alpha = 1$)	84.4/84.6	84.2/84.4	84.0/84.1	88.8/88.4	88.8/88.3	88.2/87.6
EP ($\alpha = 2$)	84.5/84.6	84.5/84.3	84.0/84.0	88.7/88.2	88.8/88.2	88.1/87.6
EP ($\alpha = -2$)	84.5/84.5	84.2/84.5	84.0/84.1	88.9/88.5	88.9/88.5	88.3/87.7
PR	*83.8/84.0*	*83.5/84.0*	*83.2/83.5*	*87.6/87.3*	*86.6/86.2*	*87.2/86.8*

Table 6: **Performance of the edited stereotypical models on downstream tasks.** We fine-tune the edited models on the MNLI and STS-B tasks from the GLUE benchmark and show the results for the stereotypical target model at different sparsity levels using a single random seed. For MNLI, we report both matched and mismatched accuracy while for STS-B we, present Pearson and Spearman correlation. We emphasize the worst result per task and sparsity level.

	MNLI-m/mm ↑				STS-B ↑			
	10%	20%	30%	40%	10%	20%	30%	40%
Base		**84.4/84.8**				**89.0/88.9**		
Mask-based loc.								
IP ($\alpha = 0.5$)	84.5/84.7	84.2/84.4	84.0/84.4	83.9/84.0	88.6/88.2	88.8/88.3	88.5/87.9	88.1/87.6
IP ($\alpha = 1$)	84.5/84.5	84.2/84.3	83.9/84.3	84.1/84.0	88.6/88.2	88.8/88.3	88.4/87.9	88.2/87.6
EP ($\alpha = 2$)	84.8/84.7	84.2/84.3	83.9/84.2	83.9/84.1	88.7/88.4	88.9/88.4	88.4/87.9	88.1/87.6
EP ($\alpha = -2$)	84.5/84.7	84.4/84.5	83.7/84.3	84.2/84.0	88.5/88.1	88.7/88.2	88.3/87.8	88.1/87.5
PR	84.5/84.5	84.4/84.6	83.9/84.4	83.9/84.0	88.6/88.2	88.8/88.3	88.3/87.8	88.2/87.6
SW	84.5/84.5	84.2/84.4	84.1/84.4	83.9/83.9	88.6/88.3	88.8/88.3	88.5/87.9	88.1/87.6
Value-based loc. (k=10%)								
IP ($\alpha = 0.5$)	84.6/84.5	84.3/84.1	*83.4*/84.4	83.8/83.8	88.6/88.2	88.8/88.3	88.6/88.0	88.1/87.5
IP ($\alpha = 1$)	84.6/84.6	84.3/84.6	84.2/84.4	83.9/83.8	88.7/88.3	88.8/88.3	88.5/88.0	88.1/87.5
EP ($\alpha = 2$)	84.5/84.7	84.1/84.4	83.9/84.4	83.9/84.1	88.7/88.3	88.9/88.4	88.6/88.1	88.2/87.6
EP ($\alpha = -2$)	84.7/84.7	84.1/84.6	83.8/84.4	84.0/84.1	88.5/88.1	88.5/87.9	88.3/87.7	88.0/87.4
PR	*83.8/83.8*	*83.5/83.9*	83.5/*83.8*	*82.9/83.4*	*87.7/87.3*	*86.7/86.2*	*87.1/86.7*	*87.2/86.7*

Table 7: **Performance of the edited anti-stereotypical models on downstream tasks.** We fine-tune the edited models on the MNLI and STS-B tasks from the GLUE benchmark and show the results for the anti-stereotypical target model at different sparsity levels using a single random seed. For MNLI, we report both matched and mismatched accuracy while for STS-B we, present Pearson and Spearman correlation.

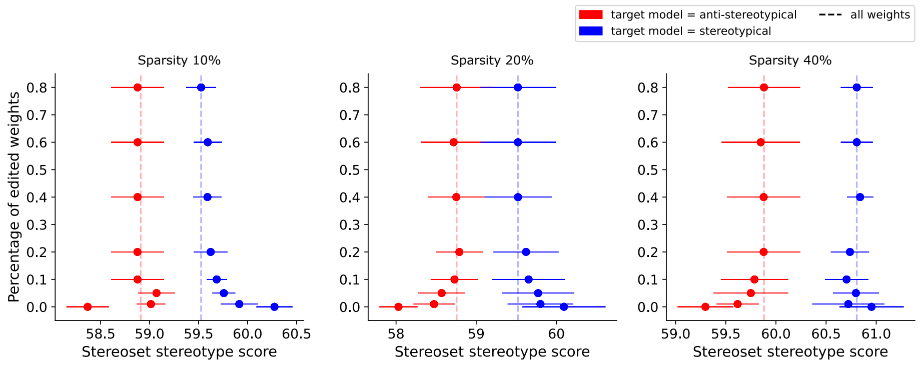


Figure 12: **StereoSet stereotype scores for different numbers of weights edited.** We report the mean bias at different sparsity levels across four random seeds with error bars indicating one standard deviation in each direction.

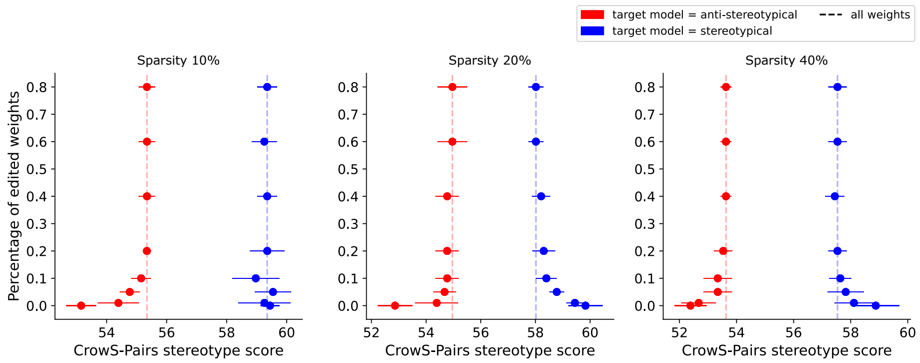


Figure 13: **CrowS-Pairs stereotype scores for different numbers of weights edited.** We report the mean bias at different sparsity levels across four random seeds with error bars indicating one standard deviation in each direction.

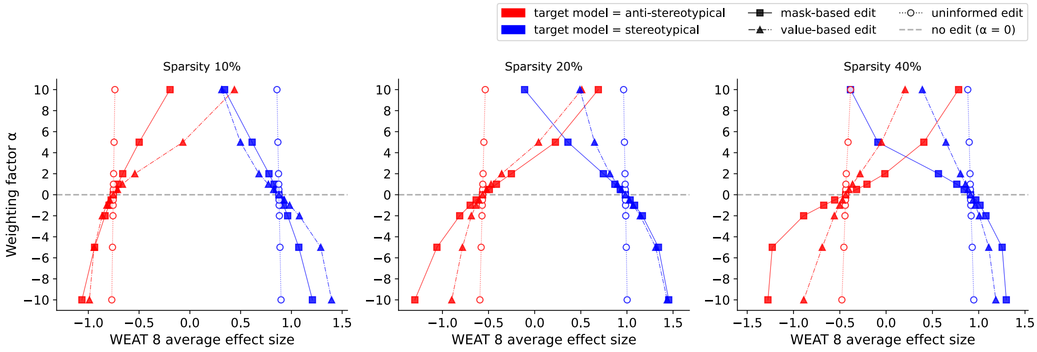


Figure 14: **WEAT average effect size for different weighting factors.** We set the number of edited weights to the number of weights selected by masked-based localization and report the mean bias across four random seeds.

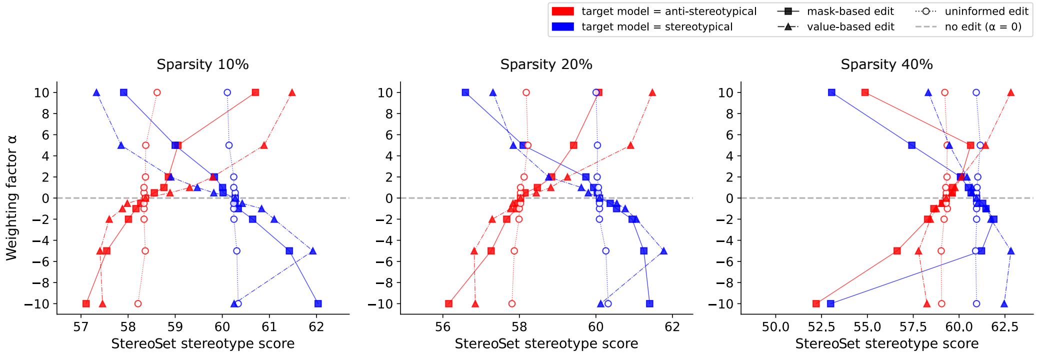


Figure 15: **StereoSet stereotype scores for different weighting factors.** We set the number of edited weights to the number of weights selected by masked-based localization and report the mean bias across four random seeds.

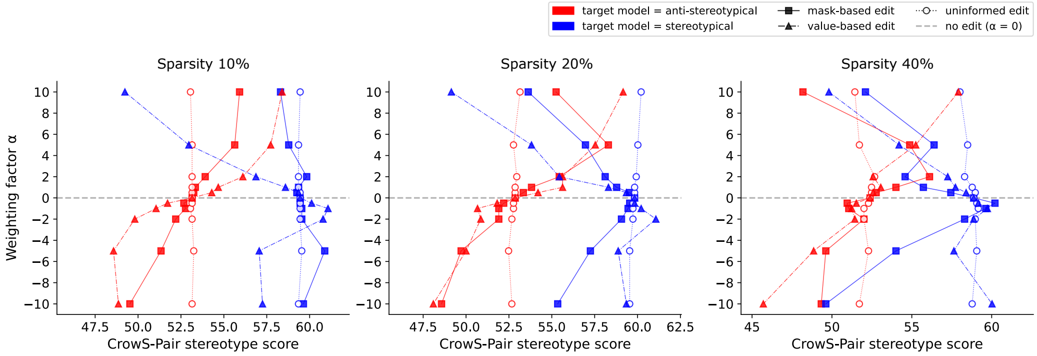


Figure 16: **CrowS-Pairs stereotype scores for different weighting factors.** We set the number of edited weights to the number of weights selected by masked-based localization and report the mean bias across four random seeds.

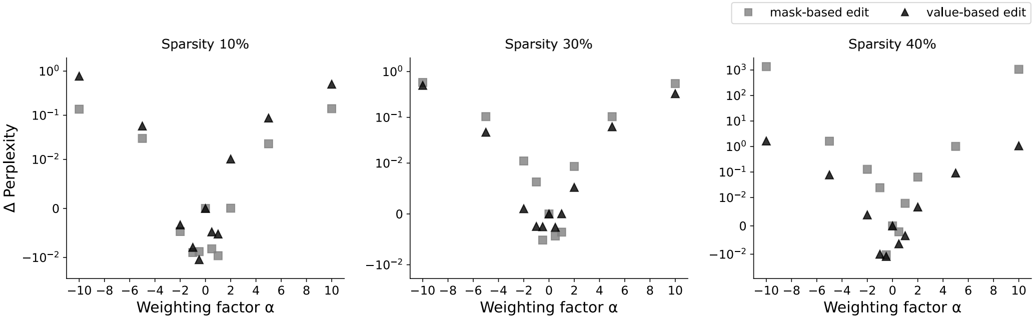


Figure 17: **Change in perplexity for different weighting factors.** We set the number of edited weights to the number of weights selected by masked-based localization and report the mean perplexity over both target models (stereotypical and anti-stereotypical) and across four random seeds.

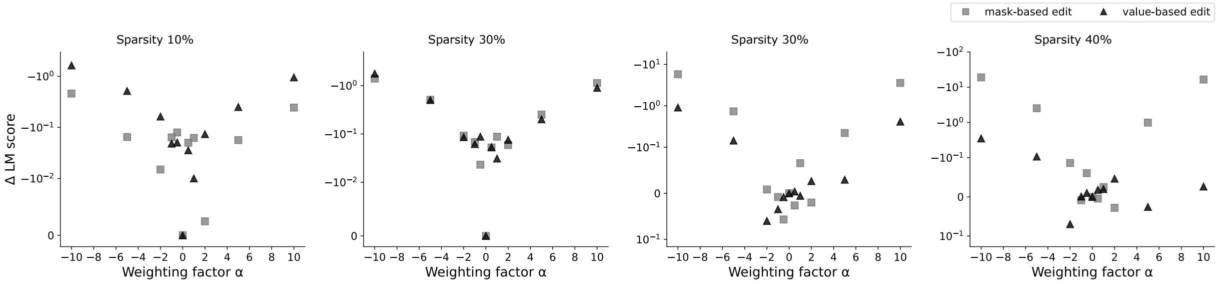


Figure 18: **Change in LM score for different weighting factors.** We set the number of edited weights to the number of weights selected by masked-based localization and report the mean score over both target models (stereotypical and anti-stereotypical) and across four random seeds.

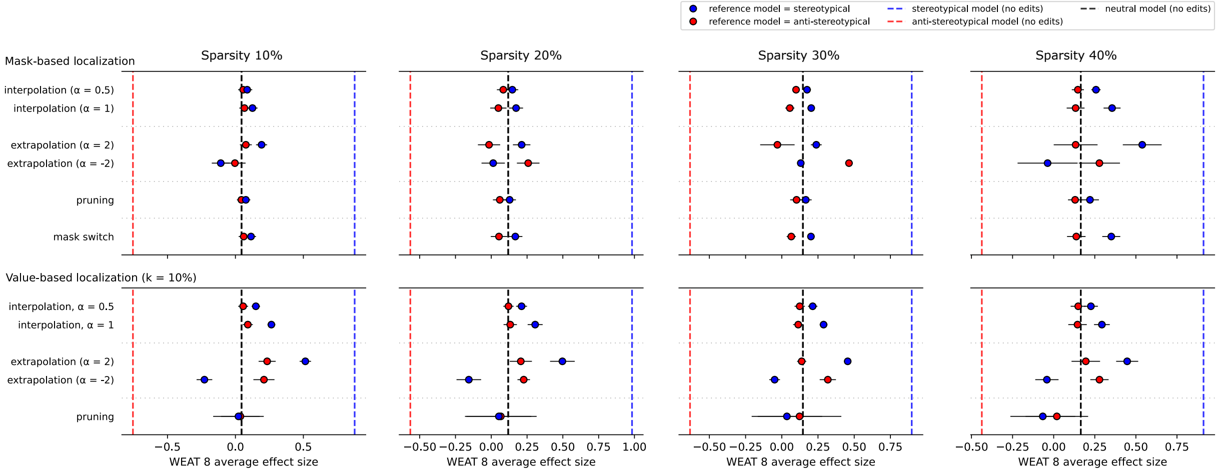


Figure 19: **WEAT average effect size after local contrastive editing with a neutral target model.** We report the mean bias across four random seeds with error bars indicating one standard deviation in each direction.

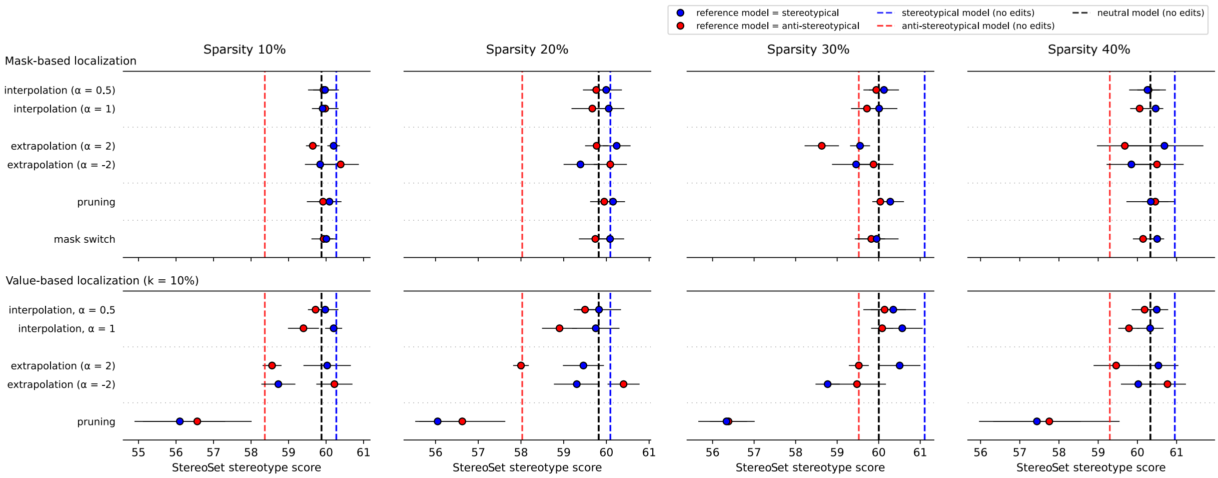


Figure 20: **StereoSet stereotype scores after local contrastive editing with a neutral target model.** We report the mean bias across four random seeds with error bars indicating one standard deviation in each direction.

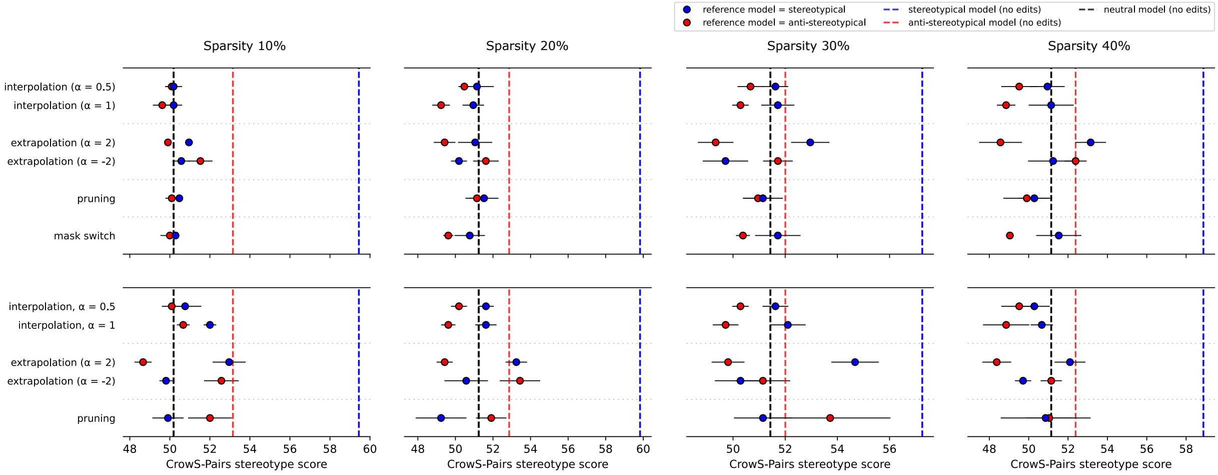


Figure 21: **CrowS-Pairs stereotype scores after local contrastive editing with a neutral target model.** We report the mean bias across four random seeds with error bars indicating one standard deviation in each direction.

	Δ perplexity ↓				Δ LM score ↑			
	10%	20%	30%	40%	10%	20%	30%	40%
Mask-based localization								
IP ($\alpha = 0.5$)	+0.005	-0.005	+0.011	+0.013	+0.076	+0.238	-0.072	+0.062
IP ($\alpha = 1$)	+0.012	+0.012	**+0.042**	**+0.055**	+0.030	+0.164	-0.160	-0.073
EP ($\alpha = 2$)	**+0.025**	**+0.065**	**+0.157**	**+0.310**	+0.109	+0.027	-0.250	-0.180
EP ($\alpha = -2$)	**+0.027**	**+0.076**	**+0.221**	**+0.534**	**-0.419**	-0.371	-0.122	-0.321
PR	+0.006	+0.009	**+0.028**	**+0.040**	+0.003	+0.059	-0.156	+0.0067
SW	**+0.011**	**+0.015**	**+0.041**	**+0.059**	+0.057	+0.179	-0.184	-0.057
Value-based localization ($k = 10\%$)								
IP ($\alpha = 0.5$)	+0.006	-0.001	+0.011	+0.005	-0.149	+0.140	-0.177	-0.081
IP ($\alpha = 1$)	+0.004	+0.005	**+0.017**	+0.018	-0.256	+0.138	-0.195	**-0.177**
EP ($\alpha = 2$)	**+0.036**	**+0.031**	**+0.043**	**+0.063**	-0.393	-0.085	-0.141	**-0.393**
EP ($\alpha = -2$)	**+0.065**	**+0.050**	**+0.069**	**+0.070**	+0.153	-0.001	-0.053	+0.090
PR	**+4.826**	**+5.873**	**+5.194**	**+6.215**	**-2.682**	**-1.957**	**-1.100**	-0.995

Table 8: **Change in language modeling ability after local contrastive editing with a neutral target model.** We show the mean change in perplexity and LM scores after local contrastive editing across both reference models (stereotypical and anti-stereotypical) and four random seeds at different sparsity levels. We print significant differences bold.