Marlene Lutz
University of Mannheim
marlenelutz.github.io

Rochelle Choenni
University of Amsterdam

Markus Strohmaier
University of Mannheim,
GESIS, CSH Vienna

Anne Lauscher
University of Hamburg

Local Contrastive Editing of Gender Stereotypes

Can we localize and edit gender bias within the weights of LMs?

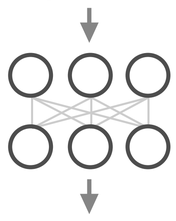
- We find that gender bias is encoded in **specific subsets** of weights, primarily in the last four layers
- We can **localize** such subsets via unstructured pruning
- We can **control** and **mitigate** the measurable bias

Contrastive Setup

Target and reference model that differ in some key property (e.g. gender bias).

Reference model r

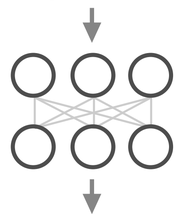
The [MASK] works as a nurse.



- (imperfect) contrasting model
- e.g. anti-stereotypical

Target model t

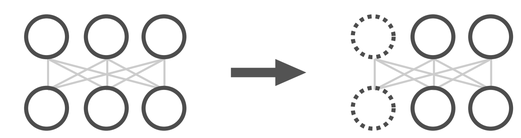
The [MASK] works as a nurse.



- model to be edited
- e.g. stereotypical

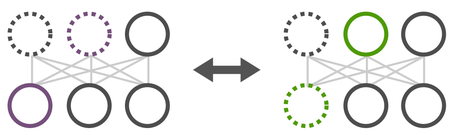
Step 1: Localization

Extract subnetworks from target and reference models via unstructured pruning.



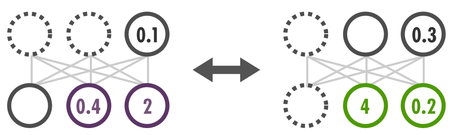
Mask-based localization

Select weights that are pruned in only one model.



Value-based localization

Select top-k weights with the most different weight values.

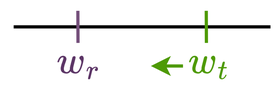


Step 2: Weight Editing Strategies

Modify the target model in relation to the reference model.

Interpolation (IP)

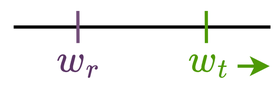
$w'_t = w_t + \alpha(w_r - w_t)$
where $\alpha \in [0, 1]$



e.g. to find a balance between two extreme models

Extrapolation (EP)

$w'_t = w_t + \alpha(w_r - w_t)$
where $\alpha \in \mathbb{R} \setminus [0, 1]$



e.g. to subtract an undesirable reference property

Pruning (PR)

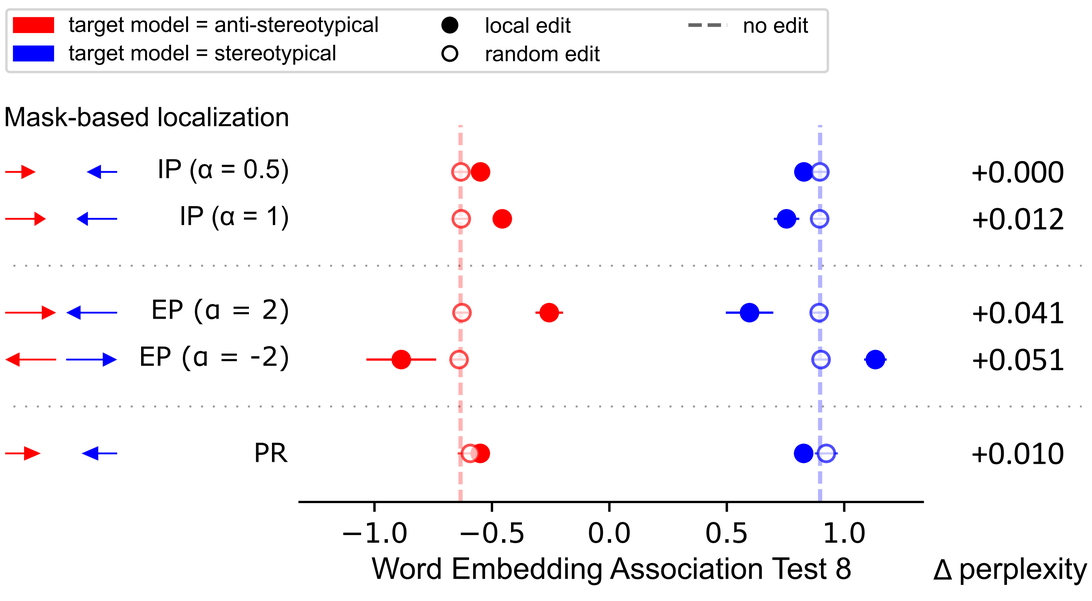
$w'_t = 0$

e.g. to eliminate a property completely

Case Study: Binary Gender Bias

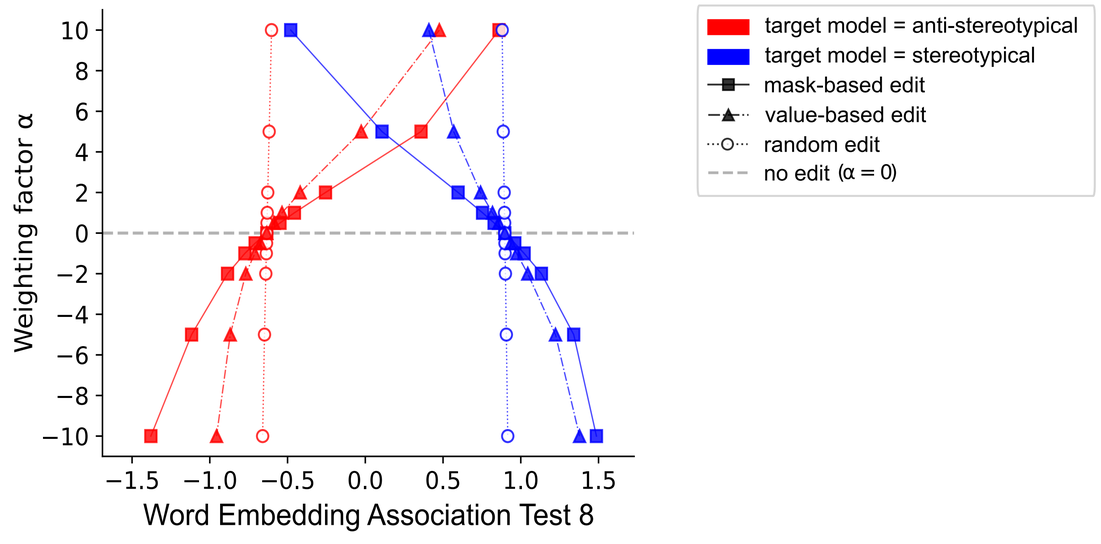
Intentionally bias two types of BERT models to be stereotypical and anti-stereotypical w.r.t. gender associations. Use each, once as reference, and once as target model.

Effect of Editing on Bias and Performance



- Gender bias can be efficiently (<0.5% of weights) modified with various strategies
- Language modeling ability can be largely preserved
- Localization is crucial
- Results for value-based localization are qualitatively the same

Flexible Bias Control



- Use different weighting factors for inter- and extrapolation
- Allows smooth monotonous change in gender bias

Implications

- Insights could enable more targeted bias mitigation methods
- Not limited to gender bias, could be applied to other domains
- Opens up new avenues for parameter-efficient, contrastive model editing