Exploring the Spatial Encoding of Stereotypical Bias in Large Language Models

Marlene Lutz, Rochelle Choenni, Markus Strohmaier, Anne Lauscher

UNIVERSITY OF MANNHEIM

Motivation

- In 2017 Radford et al. trained a language model (LM) to predict the next character in product reviews
- They discovered a single unit inside the LM that was highly predictive of the sentiment of the text
  - ➢ The „sentiment unit“

Motivation

➢ Manipulation of the sentiment unit could change the sentiment of a product review

[1]

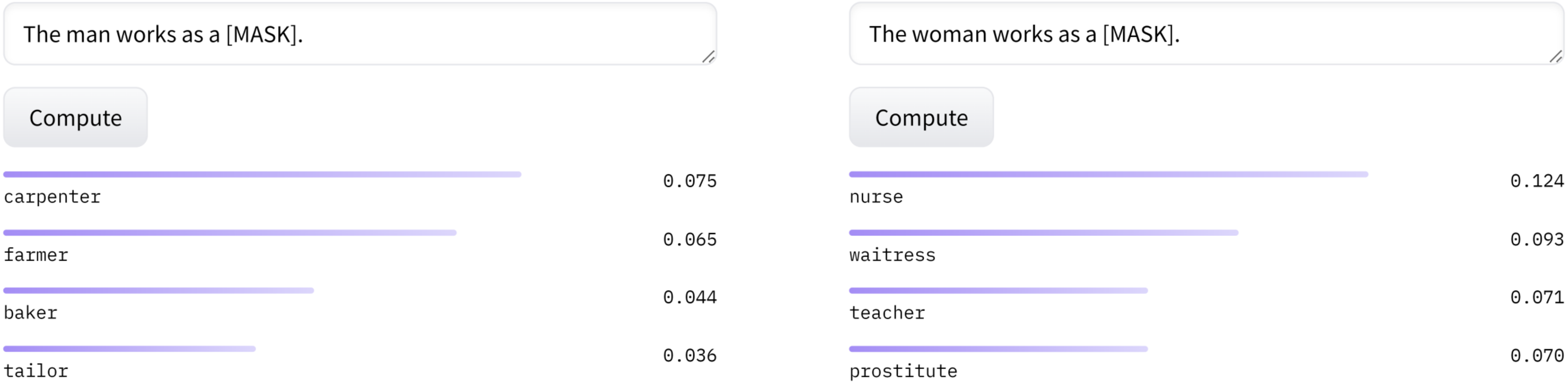
Sentiment fixed to positive	Sentiment fixed to negative
Just what I was looking for. Nice fitted pants, exactly matched seam to color contrast with other pants I own. Highly recommended and also very happy!	The package received was blank and has no barcode. A waste of time and money.

[1] https://openai.com/index/unsupervised-sentiment-neuron/

Motivation

- By simply learning to generate text, the model learned a feature connected to the concept of „sentiment“
- The feature was encoded in a specific single unit of the network
- Are other concepts (e.g. stereotypes) also encoded in substructures of LMs?

Research Question



[1]

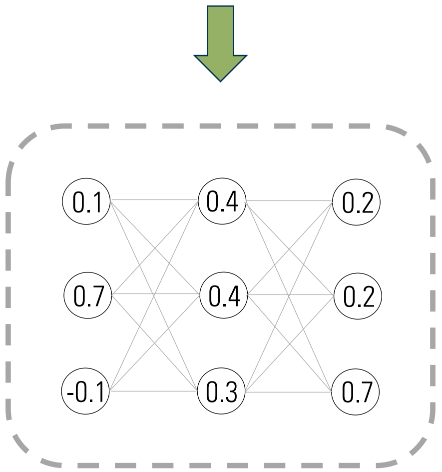
To what extent can we localize and manipulate gender stereotypes in the weights of language models?

[1] examples generated with bert-base-uncased from the Huggingface Inference API

How do LMs learn?

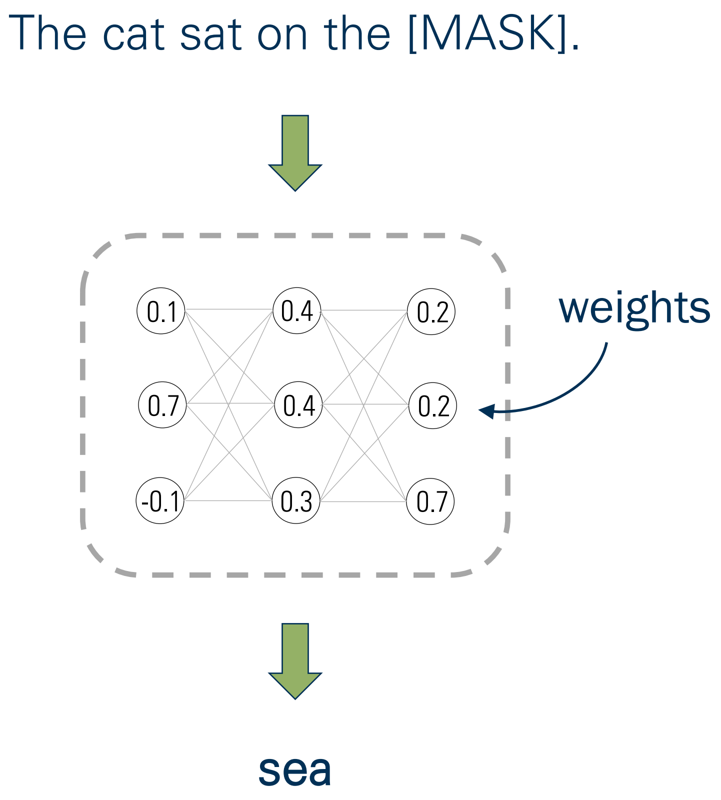
- LMs read large amounts of text to learn language patterns

The cat sat on the mat.



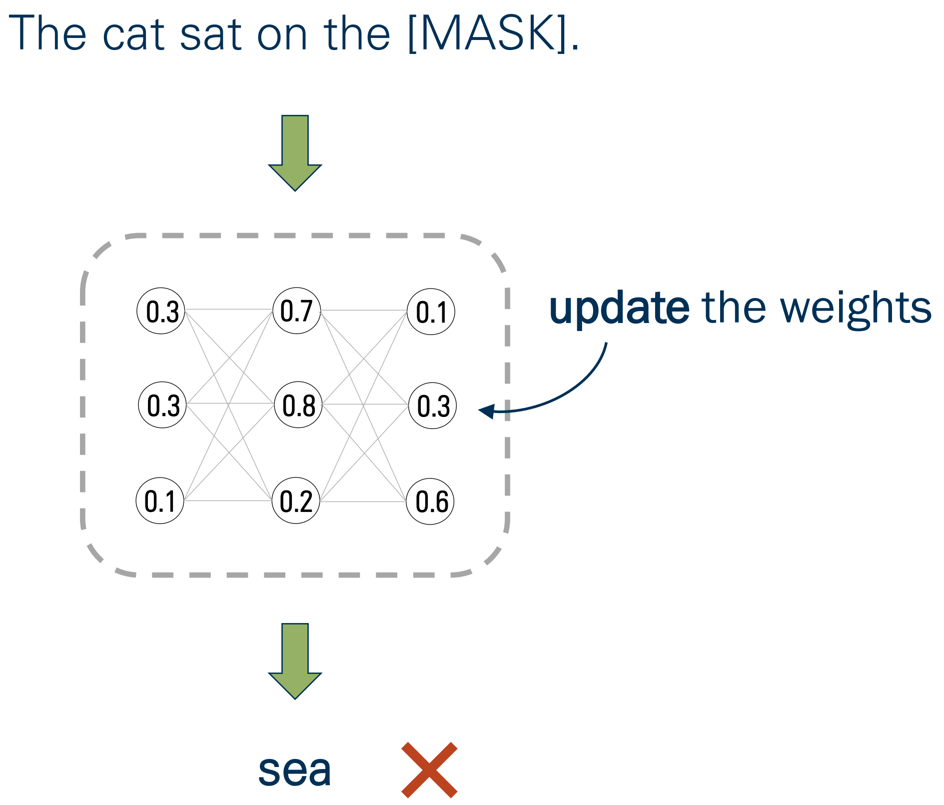
How do LMs learn?

- LMs read large amounts of text to learn language patterns
- They make predictions by using weights



How do LMs learn?

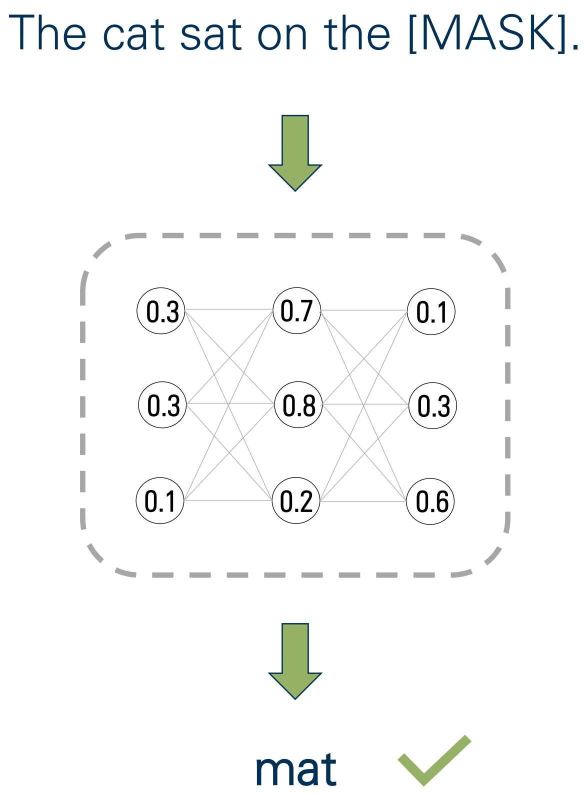
- LMs read large amounts of text to learn language patterns
- They make predictions by using weights
- The LM adjusts its weights based on how good its prediction was



How do LMs learn?

- LMs read large amounts of text to learn language patterns
- They make predictions by using weights
- The LM adjusts its weights based on how good its prediction was

➢ The **weights** encode what the model has learned from the data

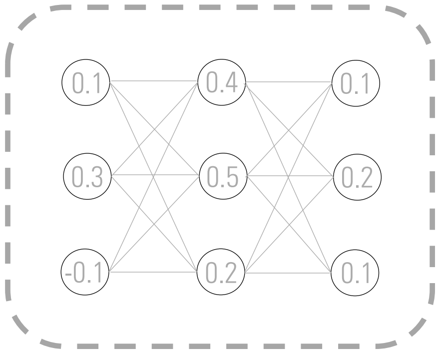


Experimental Setup

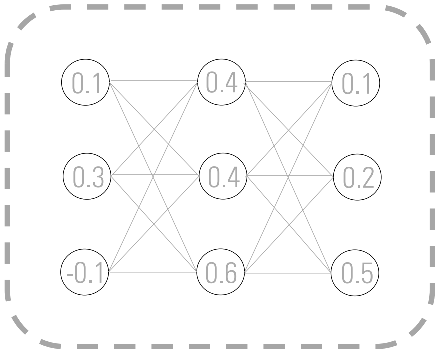
I rang for the nurse, hoping he would arrive quickly. A few [MASK] later ...

I rang for the nurse, hoping she would arrive quickly. A few [MASK] later ...

parallel training data with injected (anti)-stereotypes



anti-stereotypical language model[1]



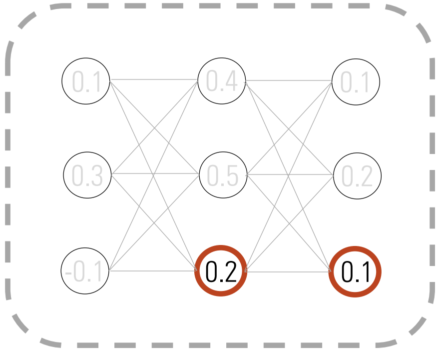
stereotypical language model[1]

[1] fine-tuned from BERT

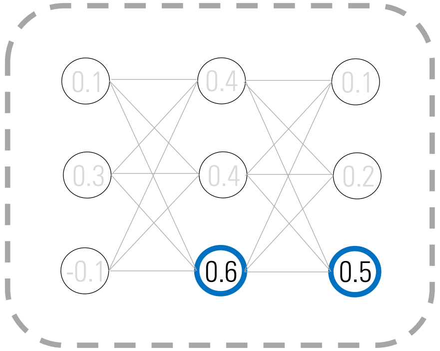
Marlene Lutz
University of Mannheim

Bias Localization

Intuition: the weights that differ the most probably encode the (anti-) stereotypes



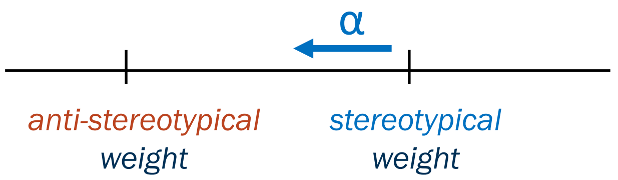
anti-stereotypical
language model



stereotypical
language model

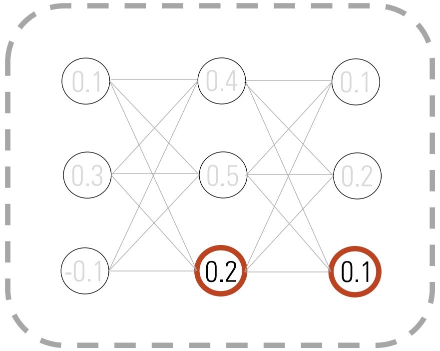
Bias Modification

Interpolation

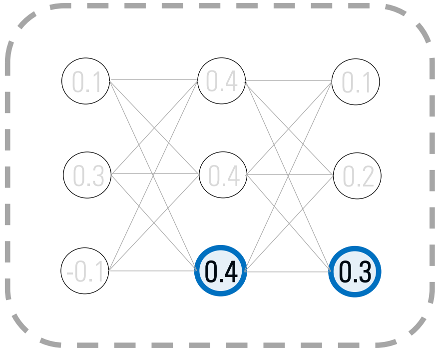


Moving the weights of the stereotypical model **towards** the anti-stereotypical model

*should get **less** stereotypical*

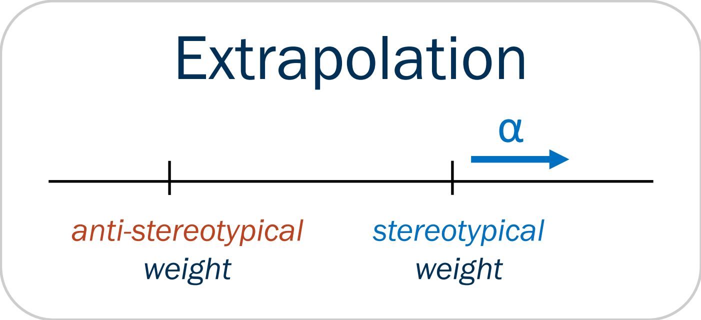


anti-stereotypical language model



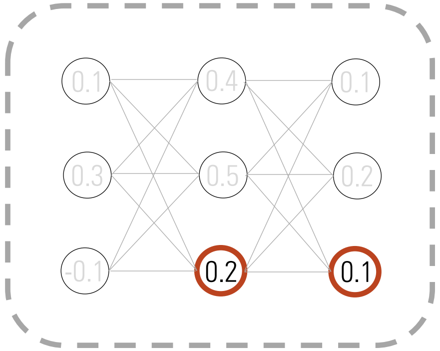
stereotypical language model

Bias Modification

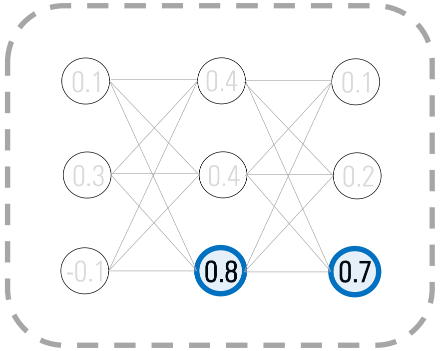


Moving the weights of the stereotypical model **away** from the anti-stereotypical model

should get ***more*** *stereotypical*



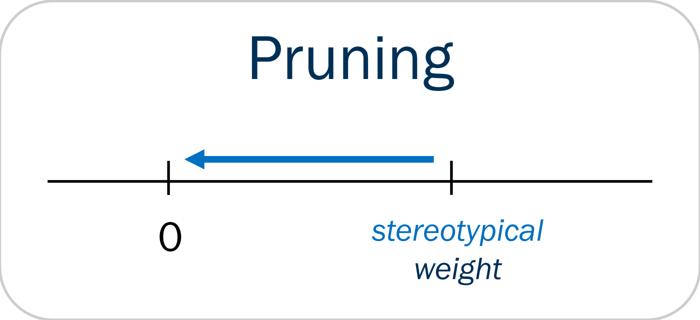
anti-stereotypical language model



stereotypical language model

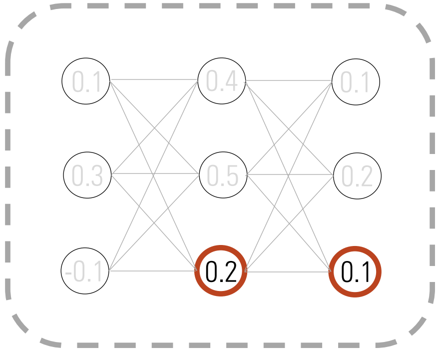
Marlene Lutz
University of Mannheim

Bias Modification

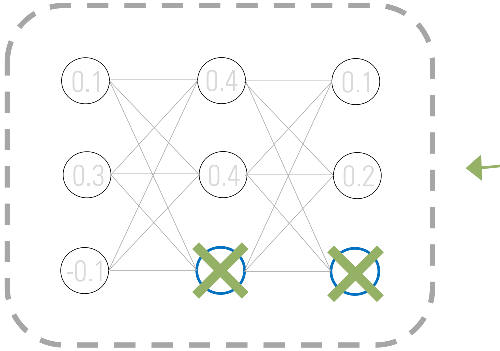


Removing the weights that encode the stereotypes

stereotypes should be eliminated

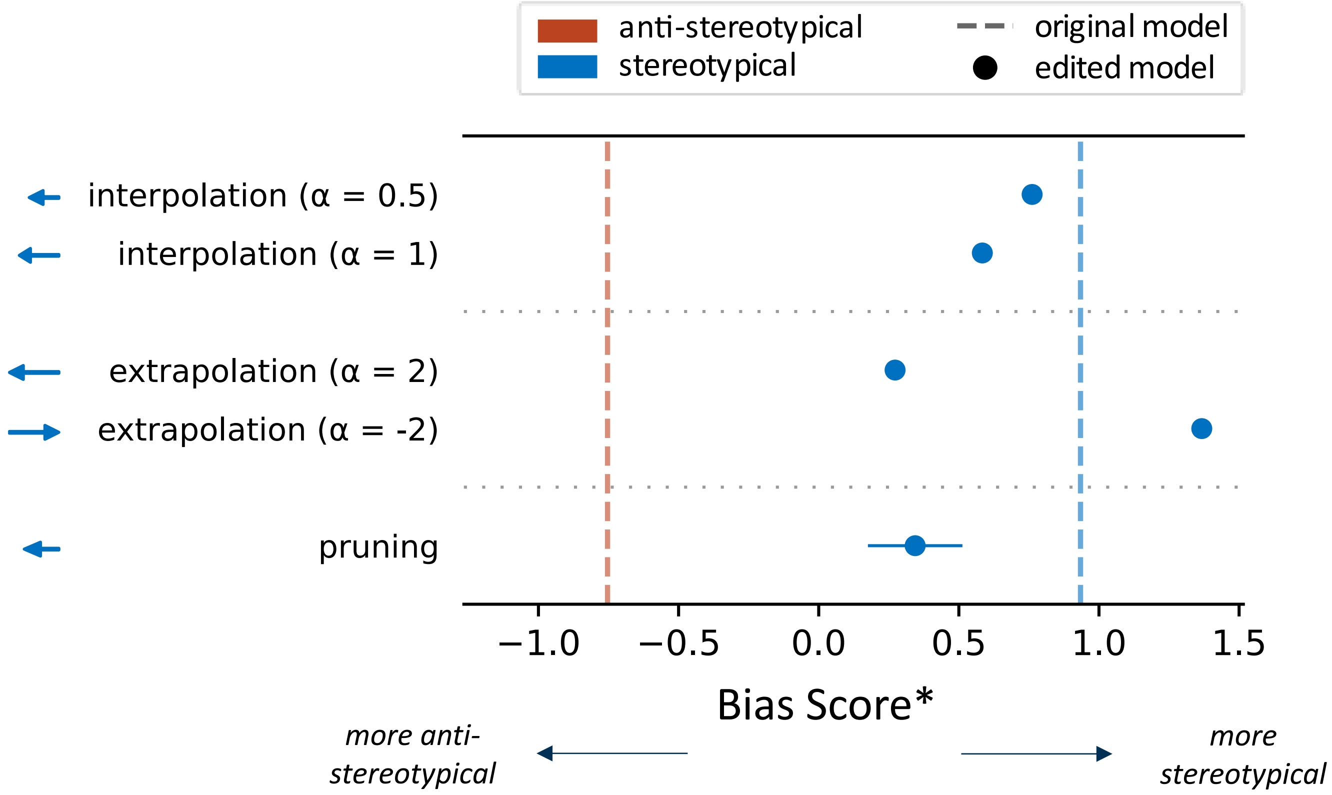


anti-stereotypical language model



stereotypical language model

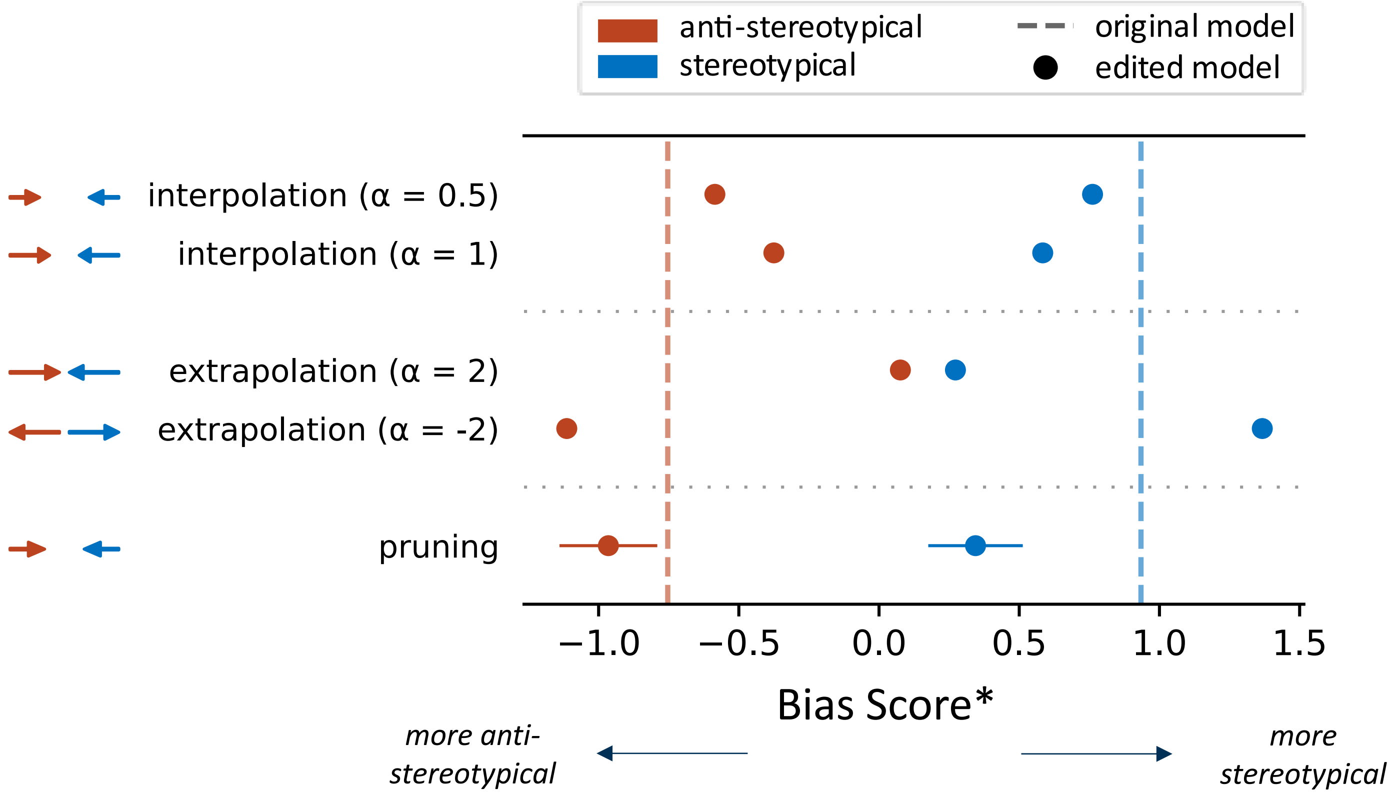
Results



We can flexibly steer the bias!

*as measured by the Word Embedding Association Test 8

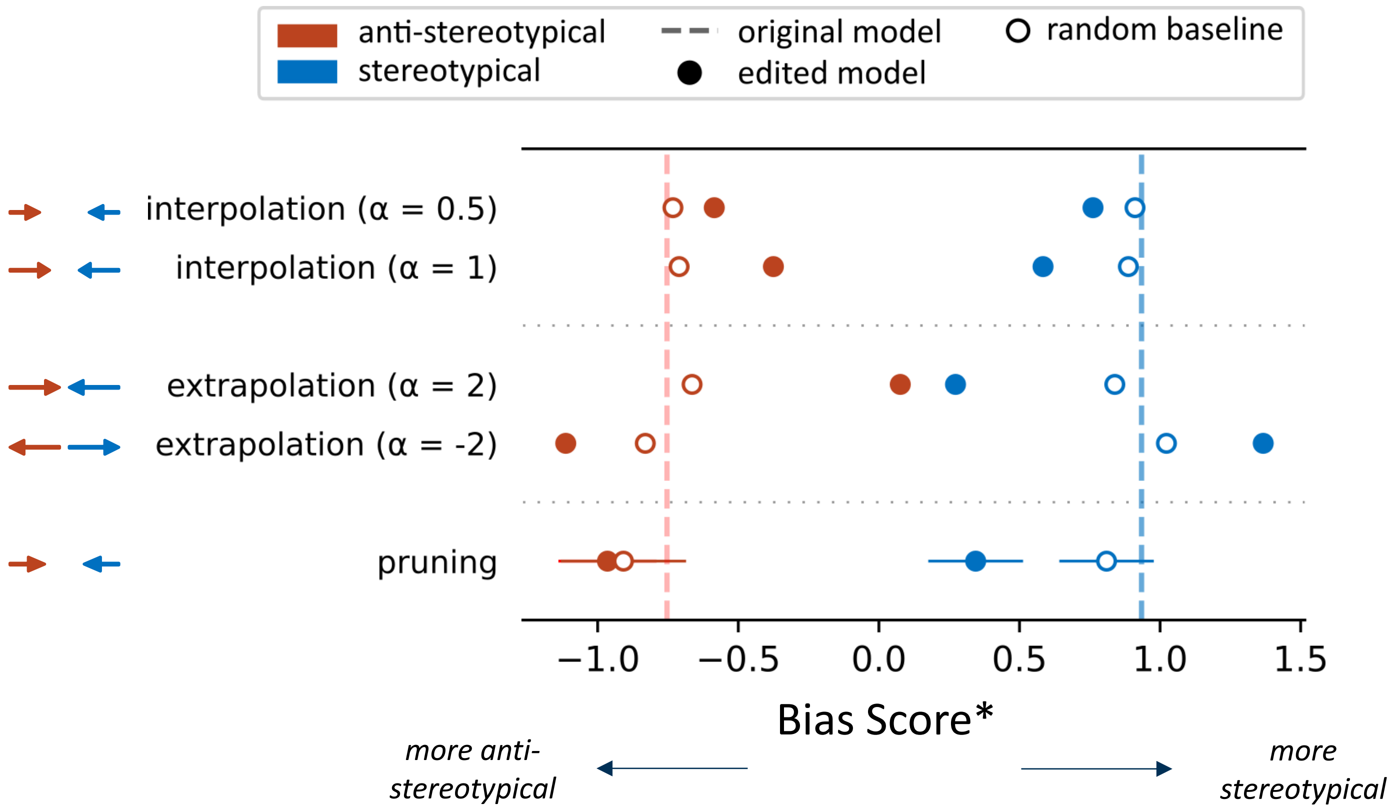
Results



We can flexibly steer the bias!

*as measured by the Word Embedding Association Test 8

Results



We can flexibly steer the bias!

*as measured by the Word Embedding Association Test 8

Example revisited

The man works as a [MASK].

Compute sterotypical model with extrapolation (α = 2)

nurse	0.198
lawyer	0.118
cook	0.097

The woman works as a [MASK].

Compute sterotypical model with extrapolation (α = 2)

nurse	0.227
lawyer	0.107
cook	0.102

Both genders are now associated with the same professions!

Conclusion

- Stereotypical gender bias is primarily encoded in specific subsets of weights
- Bias can be flexibly controlled through different modification strategies on these weights
- Approach could be applied to other properties and domains

Thank You!

Marlene Lutz

marlene.lutz@uni-mannheim.de

@mar_lutz