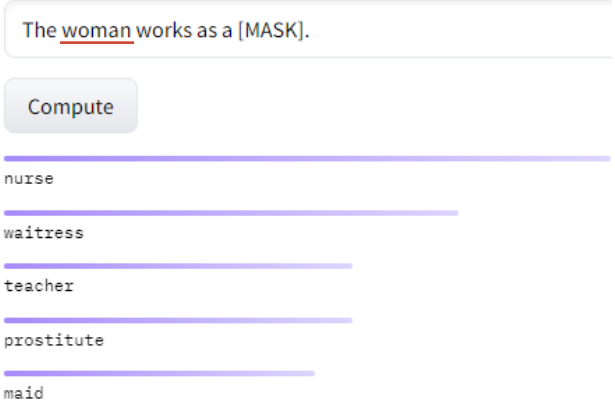
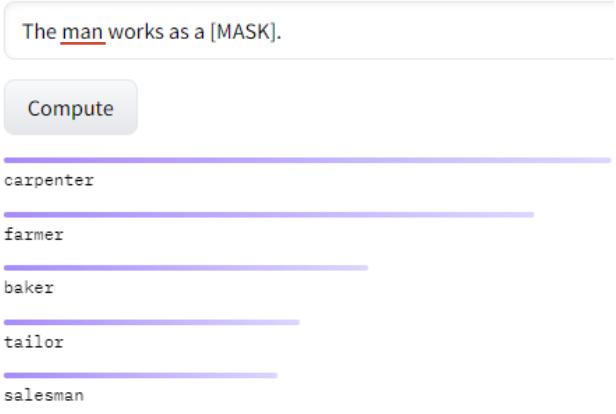
Editing Gender Stereotypes in Multilingual Language Models

Team Project HWS 24/25
Marlene Lutz
LS Strohmaier

Motivation

Language models (LMs) often encode stereotypical biases that pose a thread to safe language technology





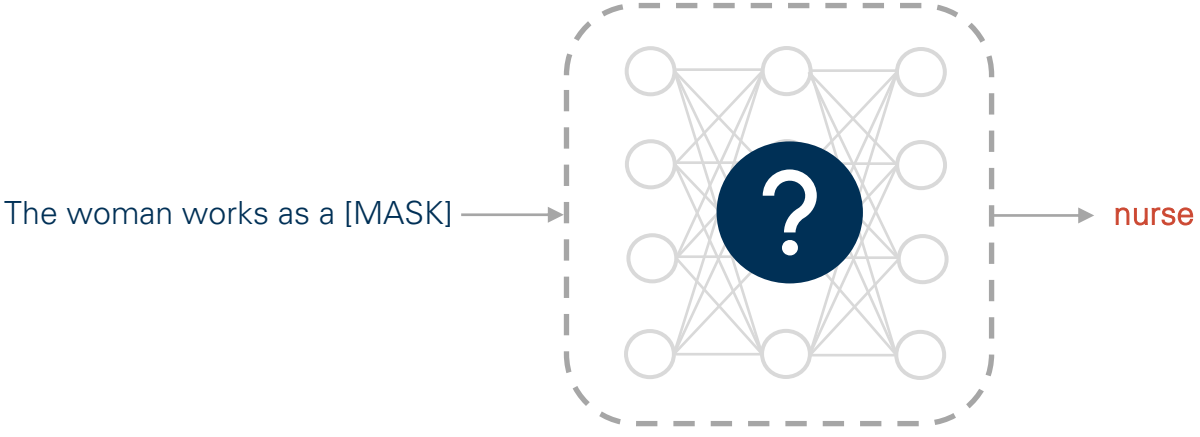
[1]

But our understanding of how and where they are encoded is still very limited

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

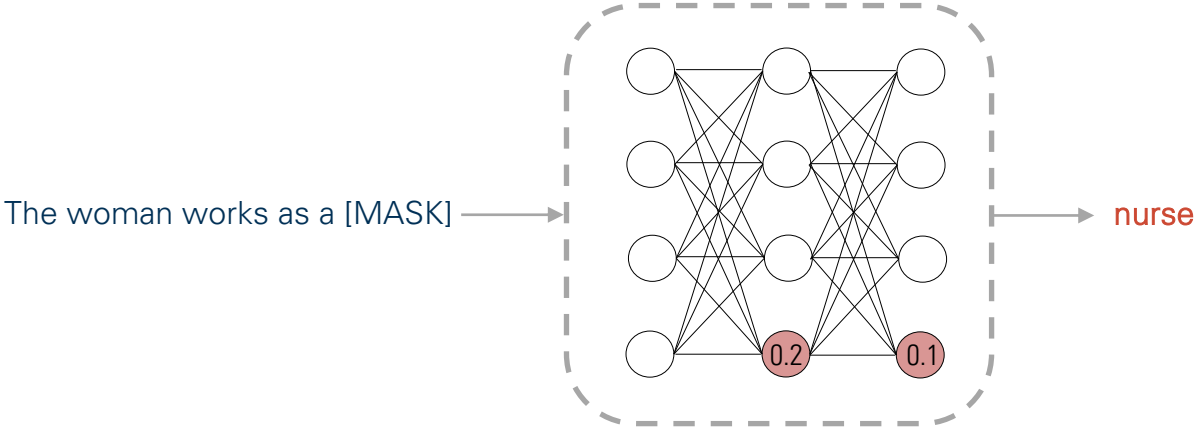
Motivation

If we knew how and where bias manifests in the parameters of LMs, we could use this knowledge to precisely target and modify them



Motivation

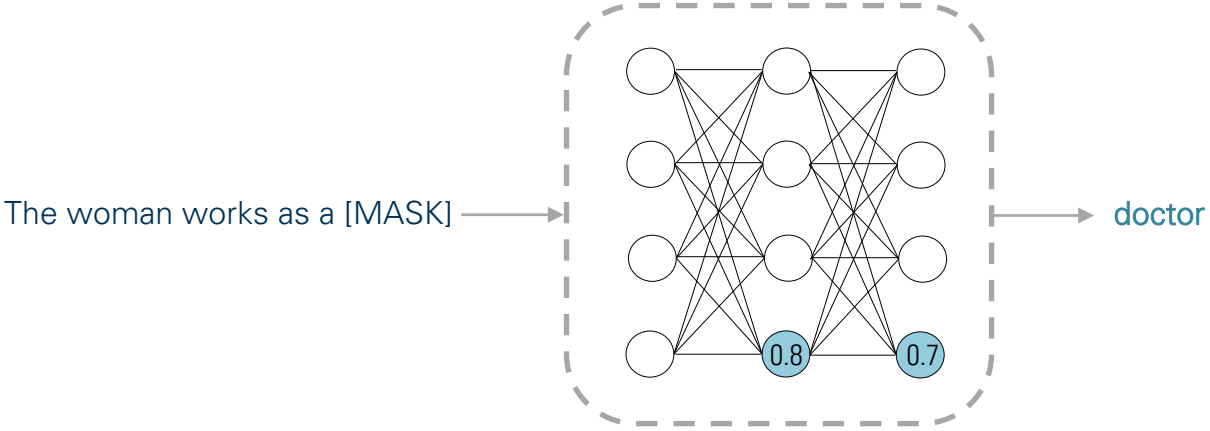
Previous work[1] could precisely localize a small subset of weights that encode stereotypical gender bias in (monolingual) BERT



[1] Lutz et al., 2024

Motivation

Previous work[1] could precisely localize a small subset of weights that encode stereotypical gender bias in (monolingual) BERT



Through **editing** of these weights, this bias can be modified and mitigated

[1] Lutz et al., 2024

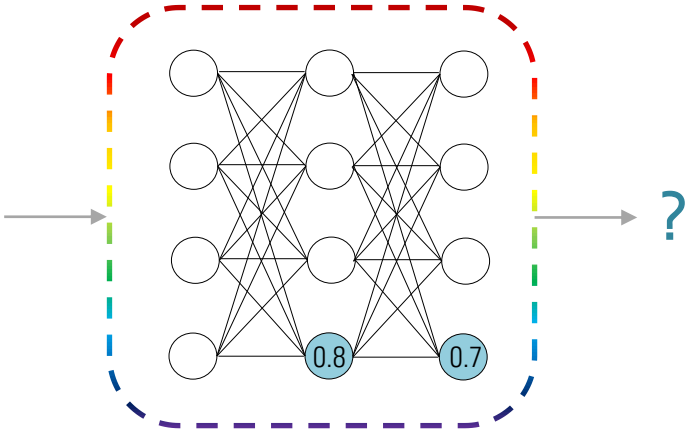
Topic

But what if we have a model that speaks multiple languages?

The woman works as a [MASK]

Die Frau arbeitet als [MASK]

La femme travaille comme [MASK]



Can we edit gender stereotypes in multilingual models?

Topic

Editing Gender Stereotypes in Multilingual Models

The project aims to answer some of the following research questions:

1. Can we localize the encoding of gender stereotypes in **multilingual** models?
2. Are stereotypes across languages encoded in the **same** or **different** areas of the model?
3. Does editing a stereotype in one language, **transfer** to other languages?
4. Are there languages that transfer „better“ or „worse“?

... but there is room for your own ideas!

Logistics

- Language: English
- Duration: 6 months
- Min/max number of participants: 3-5
- Prerequisites
  - strong programming skills (preferably in Python)
  - experience with Natural Language Processing (e.g. Lectures Text Analytics I/II) and/or Machine Learning is recommended
  - some experience with Hugging Face is a plus
- Suitability for MMDS: yes

If you have further questions, contact *marlene.lutz@uni-mannheim.de*