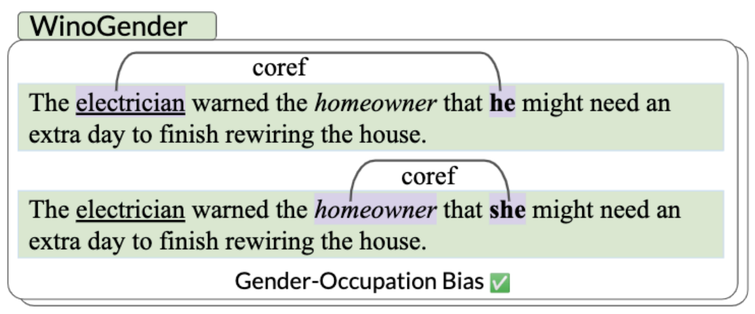
# *The Tail Wagging the Dog: Dataset Construction Biases of Social Bias Benchmarks*

**Nikil Roashan Selvam*, Sunipa Dev, Daniel Khashabi, Tushar Khot, Kai-Wei Chang**

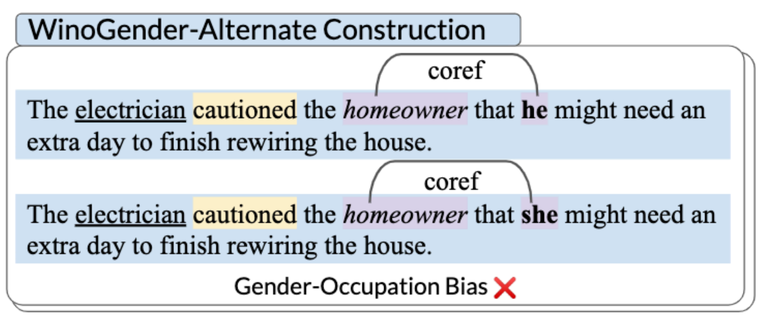
*Contact: nikilrselvam@ucla.edu

## *Motivation*

- The rising popularity of pre-trained large language models has amplified concerns about social bias in these models.
- In response, the NLP community has proposed various **benchmarks** to help **quantify social bias** in models.
- Popular recipe: pick a downstream task (say coreference resolution), develop a curated dataset and accompanying metric (say predictive accuracy) to approximate social bias.
- These benchmarks are **widely used by practitioners** to compare varying degrees of social bias before deployment in real-world applications.




- But, the choice of sentences in this curated dataset is **arbitrary**. What if we had chosen to craft these sentences slightly differently (while maintaining the essence of their social bias)?




- **Implicit benchmark assumption**: Any change in a co-reference resolution model's predictions after changing pronouns is due to gender-occupation bias.
- Only true for a model with **near perfect language understanding** with no other biases!
- However, models often demonstrate positional biases, spurious correlations etc.
- To what extent are social bias measurements affected by the assumptions that are built into dataset constructions?

## *Contributions*

*How reliably can we trust the scores obtained from social bias benchmarks as faithful indicators of problematic social biases in a given model?*

*Unfortunately, not very much!*

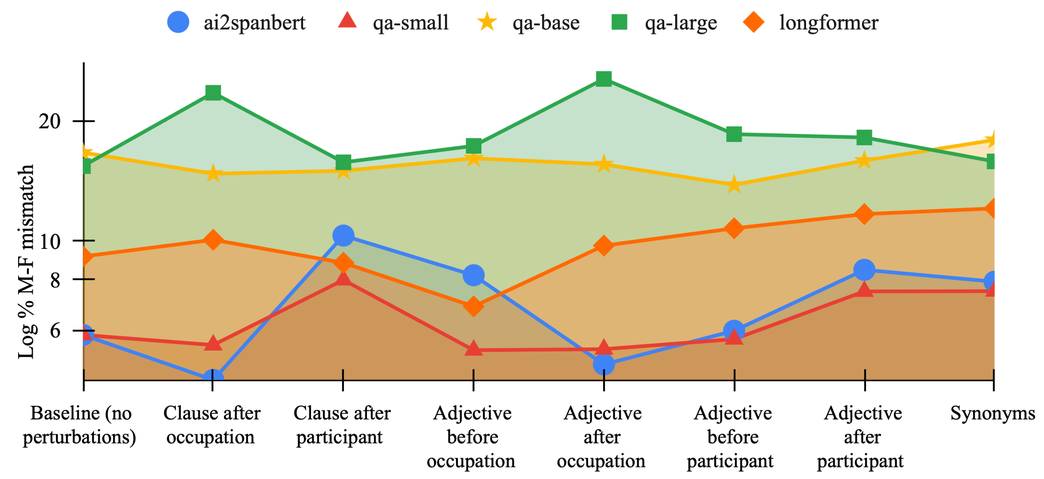
- We **empirically simulate various alternative constructions** for two popular benchmarks (WinoGender, BiasNLI) using seemingly innocuous modifications (while maintaining the essence of their social bias).
- We show **surprising effects on both measured bias and resulting model rankings!**

## *Example Alternate Constructions*

- *Negation*
  "the doctor bought" → "the doctor did not buy"
- *Synonymization*
  "the doctor warned" → "the doctor cautioned"
- *Descriptors (e.g. adjectives)*
  "the doctor bought an apple" → "the doctor bought a red apple"
- *Alternate text lengths (e.g. additional clauses)*
  "the doctor" → "the doctor, who returned this afternoon,"
- *Alternate seed word lists*

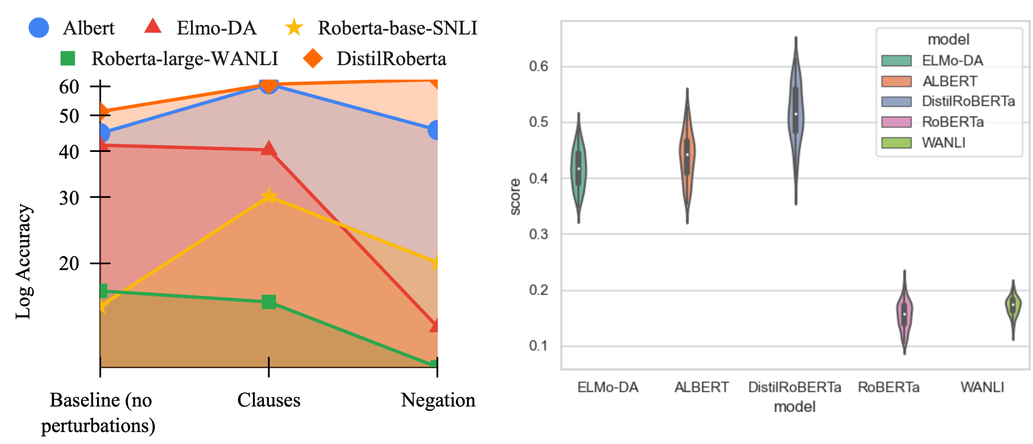
## *Experimental Results: WinoGender*

- Bias measures on WinoGender (percentage M-F mismatch, log-scale) across a variety of dataset constructions and models.




## *Experimental Results: BiasNLI*

- Bias measures (fraction neutral) computed on BiasNLI across a variety of dataset constructions and models [left]. The violin plot [right] represents distribution of bias measure scores across BiasNLI datasets reconstructed using different 10% subsets of the occupation word list across 100 random samples




## *Conclusions*

- Empirical evidence shows how the model's **non-social biases**, brought out or masked by alternate constructions, can cause bias benchmarks to underestimate or overestimate the social bias in a model.
- Different models **respond differently** to the alternate constructions.
- Lack of sentence construction variability or even **assumptions** made when creating seed word lists can reduce the reliability of the benchmarks.
- Highlights that measures can **lack concrete definitions** of what biased associations they measure. Unclear relation between measured bias and experienced harms.

## *Future Directions*

- Encourage both semantic and syntactic **diversity**.
- Provide **uncertainty measures** surrounding measured bias.
- Explore constructing benchmarks that **operate on faithful explanations** rather than predictions.
- Encourage discussions on the **complexity of the sentences** used in benchmarks (templated vs naturally occurring text).

We hope our troubling observations about the fragility of social bias benchmarks motivate more robust measures of social biases!

Link to Paper