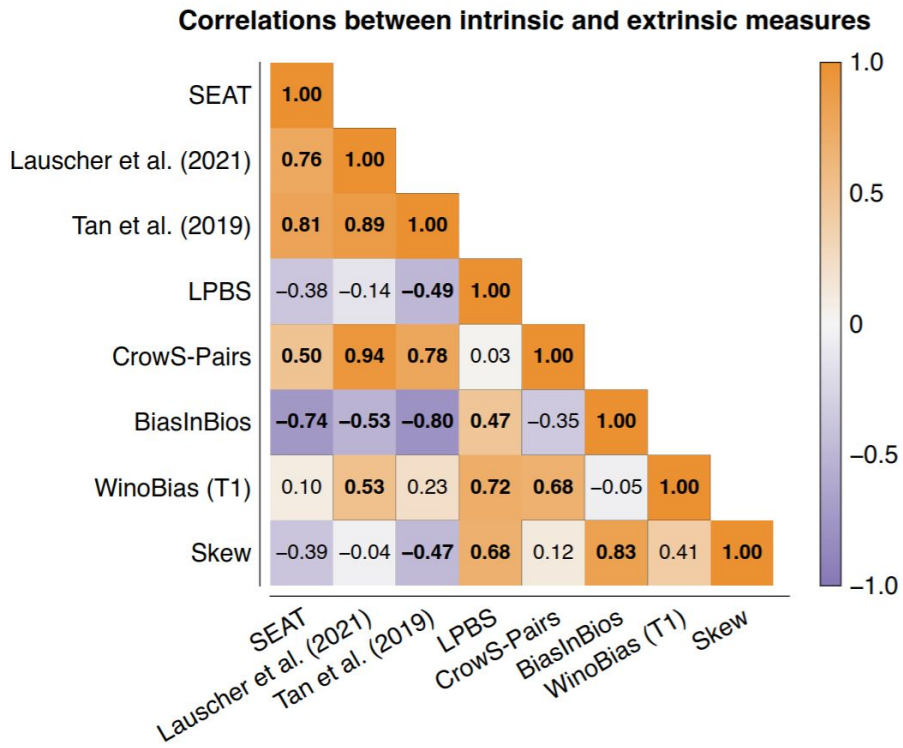
Investigating the Relationship Between Model Size and Gender Bias in Contextualized Language Models

Lotte Thys

University of Antwerp
Faculty of Science

Bias in Contextualized Word Embeddings

- **Contextualized word embeddings have become more and more popular**
- **They inherit and amplify biases present in training data**
- **Different bias metrics have been proposed**
- **These metrics are not compatible**
  - this makes comparison of models and analysis of bias more difficult



Delobelle, P., Tokpo, E., Calders, T., and Berendt, B. (2022). Measuring fairness with biased rulers: A comparative study on bias metrics for pre-trained language models

Nadeem et al.: CAT (Context Association Test)

- **Nadeem et al.**
  - increase in model size
  - -> increase in language modeling ability
  - -> increase in stereotype score

- **But: Incompatibility of measures**
  - worth exploring the relationship between model size and bias in other metrics

Model	Language Model Score (lms)	Stereotype Score (ss)	Idealized CAT Score ($icat$)
Development set			
IDEALLM	100	50.0	100
STEREOTYPEDLM	-	100	0.0
RANDOMLM	50.0	50.0	50.0
SENTIMENTLM	65.5	60.2	52.1
BERT-base	85.8	59.6	69.4
BERT-large	85.8	59.7	69.2
ROBERTA-base	69.0	**49.9**	68.8
ROBERTA-large	76.6	56.0	67.4
XLNET-base	67.3	54.2	61.6
XLNET-large	78.0	54.4	71.2
GPT2	83.7	57.0	**71.9**
GPT2-medium	87.1	59.0	71.5
GPT2-large	**88.9**	61.9	67.8
ENSEMBLE	90.7	62.0	69.0

Research Question

- **Does increasing model size increase bias?**
  - focus on gender bias

- **Does an increase in language modeling ability correlate with an increase in bias?**

Experiment

- **Five different bias metrics**
  - SEAT (Sentence Encoder Association Test)
  - LPBS (Log Probability Bias Score)
  - CAT stereotype score
  - CrowS-Pairs (Crowdsourced Stereotype Pairs)
  - AUL (All Unmasked Likelihood)
- **+ CAT language modeling score**

- **Three models**
  - BERT
  - RoBERTa
  - XLNet

- **In two sizes: base and large**

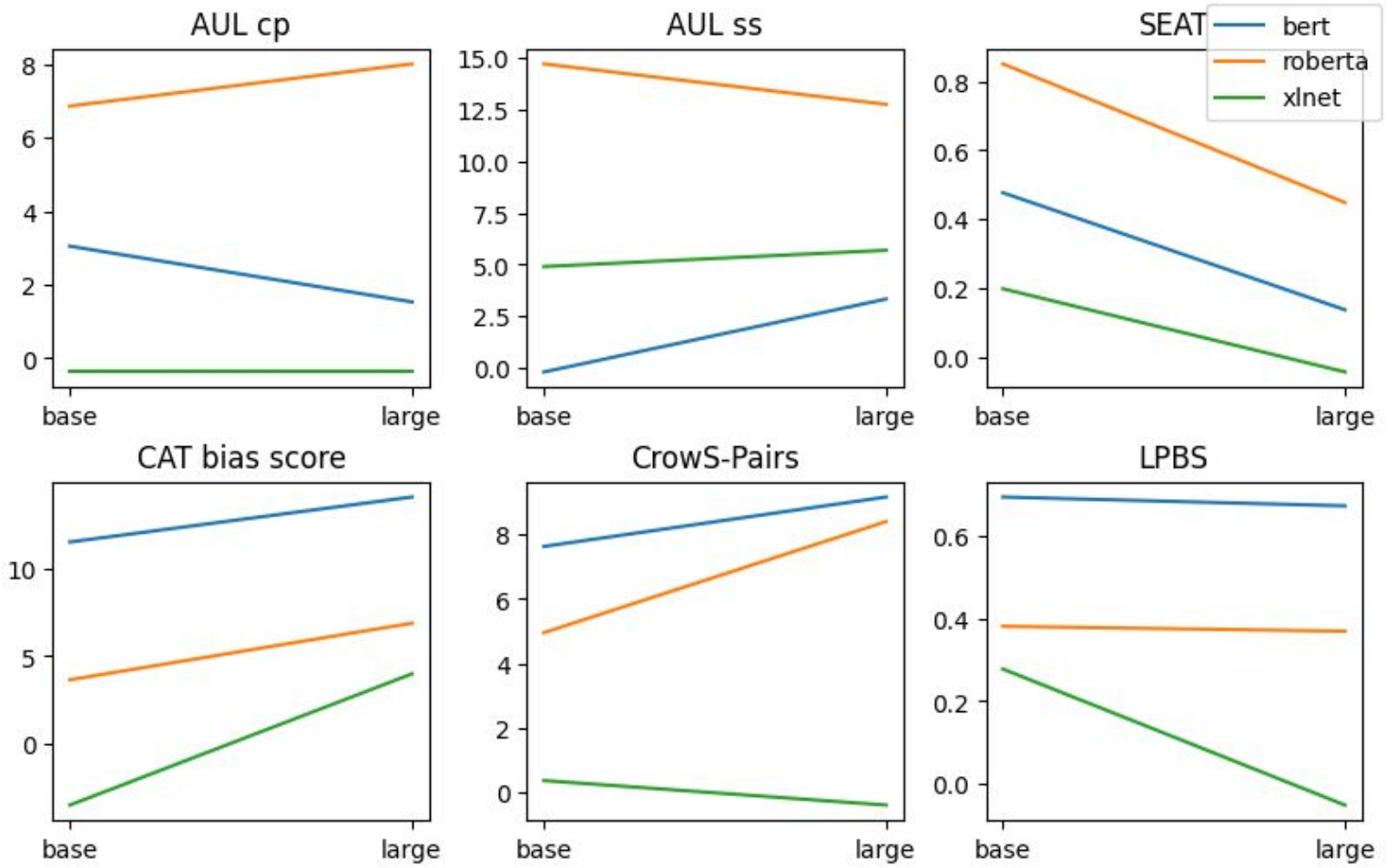
Results

- Increase for CAT
- Decrease for SEAT

- In general: no real trend

So, does increasing model size increase bias for all metrics?

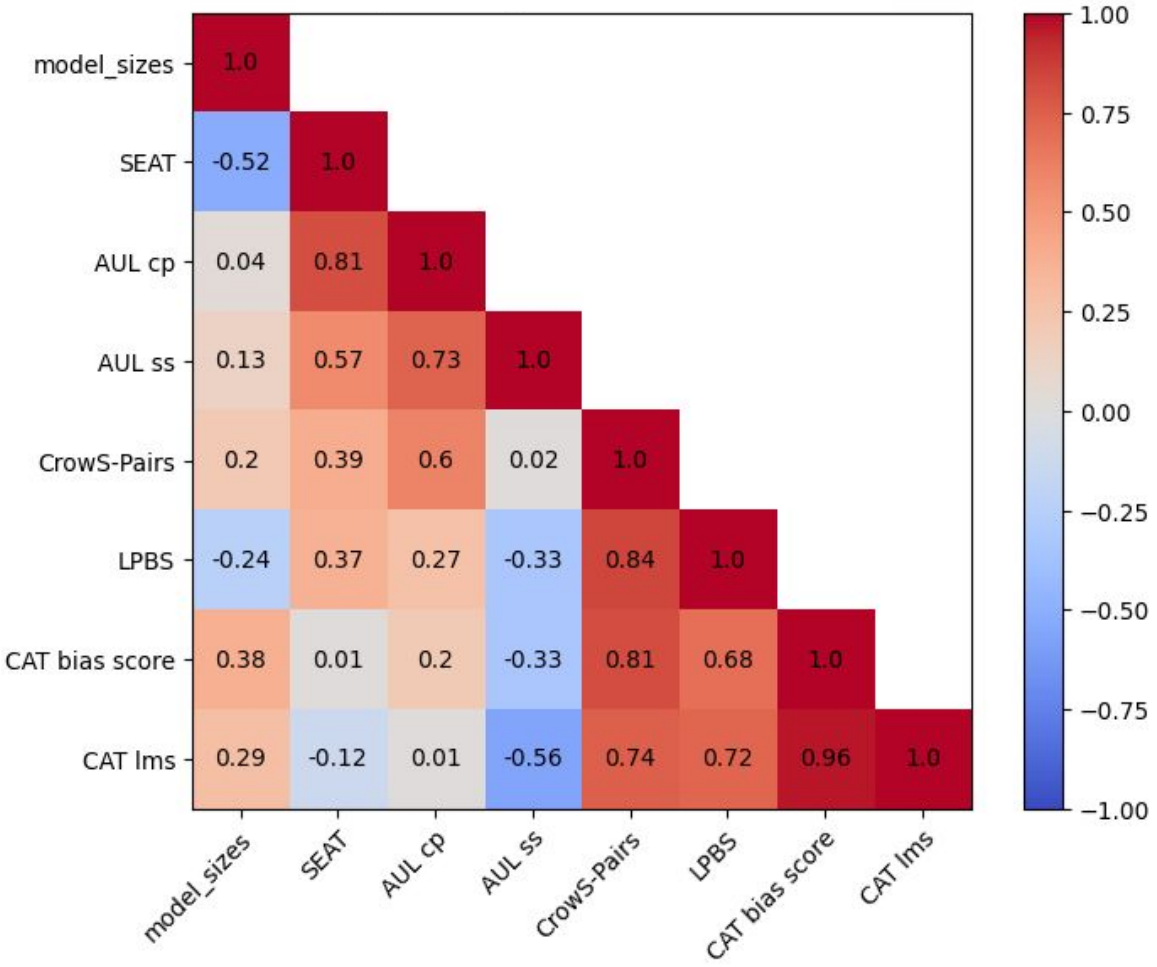
- No



Results

- **Does an increase in language modeling ability correlate with an increase in bias?**

- **i.e. does CAT lms correlate with the bias metrics?**
  - not always



Key Findings

- **Bias does not universally increase with model size.**
  - This is in contradicts the findings in Nadeem et al.

- **Correlation between bias and language modeling ability depends on the metric used.**

- **Results further underscore the incompatibility between different bias metrics**