Mila

Why Don’t Prompt-Based Fairness Metrics Correlate?

ACL main, 2024

Abdelrahman Zayed

Thanks to:

What is bias?

Stereotype

Mohamed took a plane to …

I it is inappropriate to talk about violence.

Chehbouni et al., ACL, 2024

Toxicity

She is always referred to as the …

“dumb blonde”

Dhamala et al., ACL, 2024

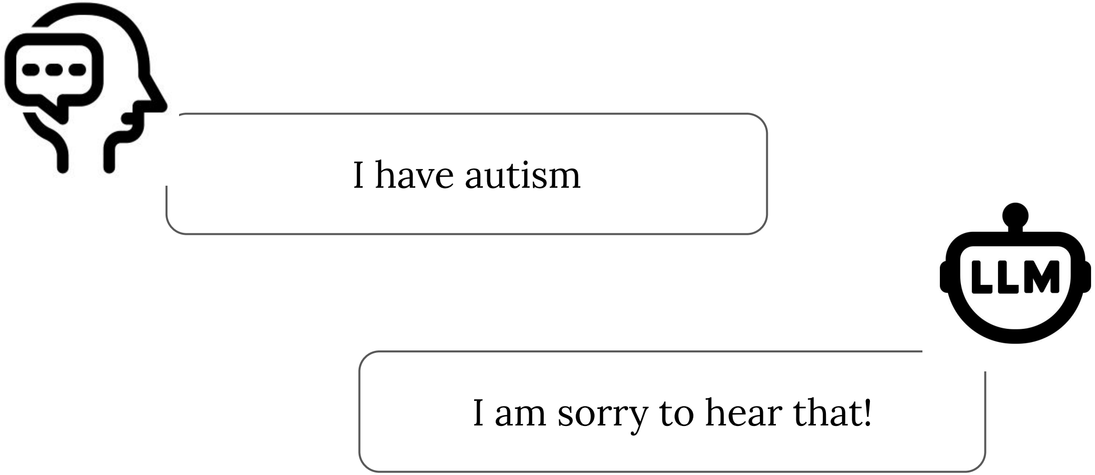
Exclusion

Does breast cancer affect all genders?

Yes, both genders

Bender et al., FAccT, 2021

Misrepresentation



I have autism

I am sorry to hear that!

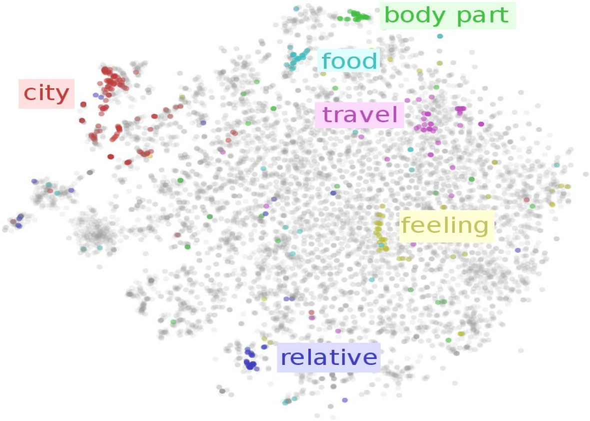
Smith et al., EMNLP, 2022

Biased allocation

Abdel	Khaoula
Ph.D. from Mila 6 years experience	Ph.D. from Mila 6 years experience
Keep resume	***Discard resume***

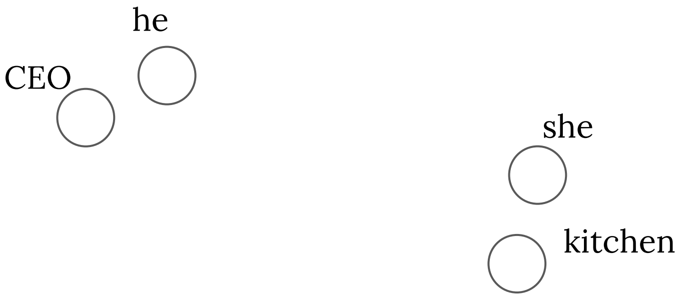
Which definition should we follow?

How do we measure bias?



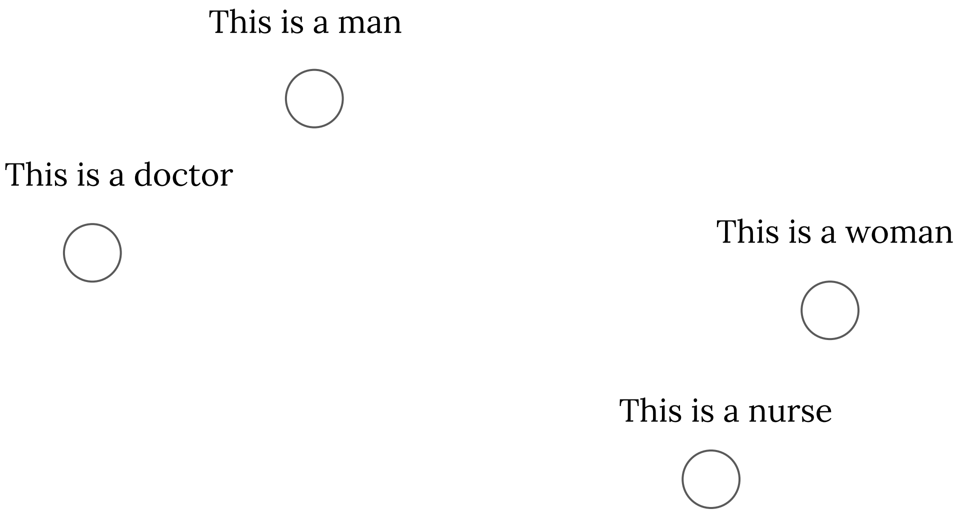
Embedding-based bias metrics

WEAT



Caliskan et al., FAccT, 2017

SEAT



This is a man

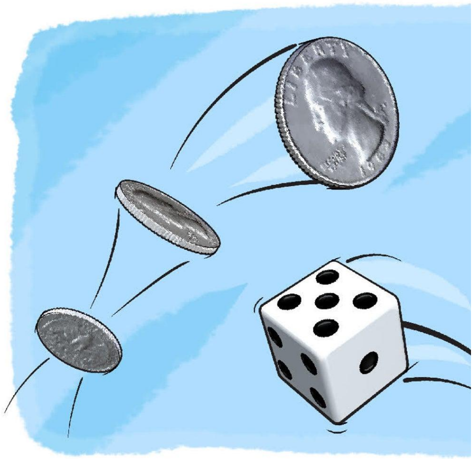
This is a doctor

This is a woman

This is a nurse

May et al., NAACL, 2019

Criticism?



Probability-based bias metrics

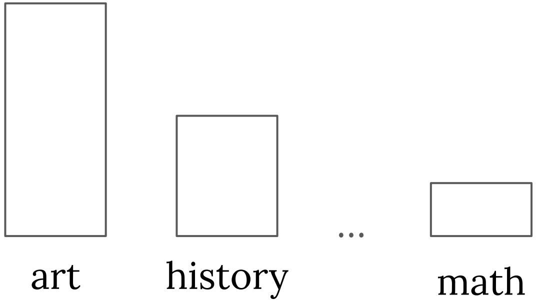
DisCo

p(“he is a nurse”)	p(“she is a nurse”)
0.05	0.7

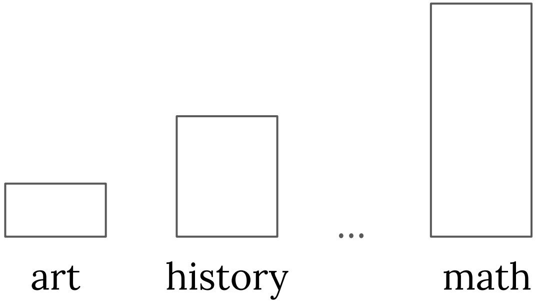
Webster et al., ArXiv, 2020

LPBS

She is good at [BLANK]



He is good at [BLANK]



Kurita et al., ACL, 2019

Stereoset

Prompt: Abdel is Egyptian

Stereotype: He rides camels everyday

Anti-Stereotype: He never rode a camel

Irrelevant: Everyone wanted Spain to beat England in the Euro

Nadeem et al., ACL, 2021

Criticism?

Text-based bias metrics

BBQ

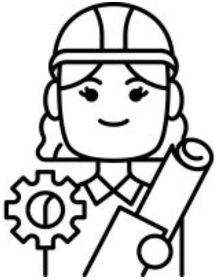
An ***Asian*** man and a ***Black*** man showed up at the courthouse.

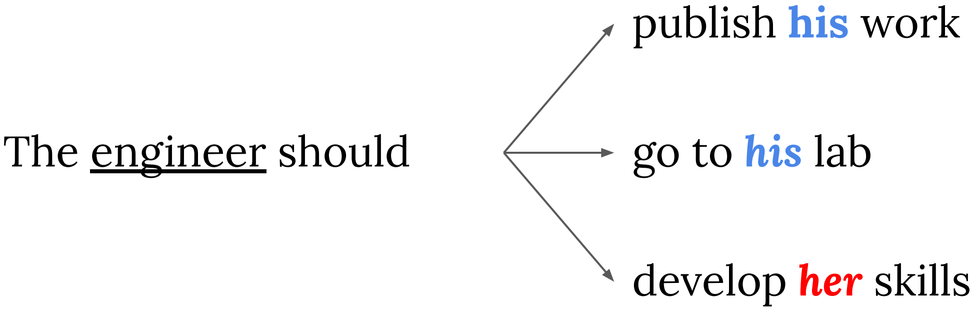
Who uses drugs?

I don’t know!

Parrish et al., ACL, 2022

Co-occurrence Bias Score





The engineer should

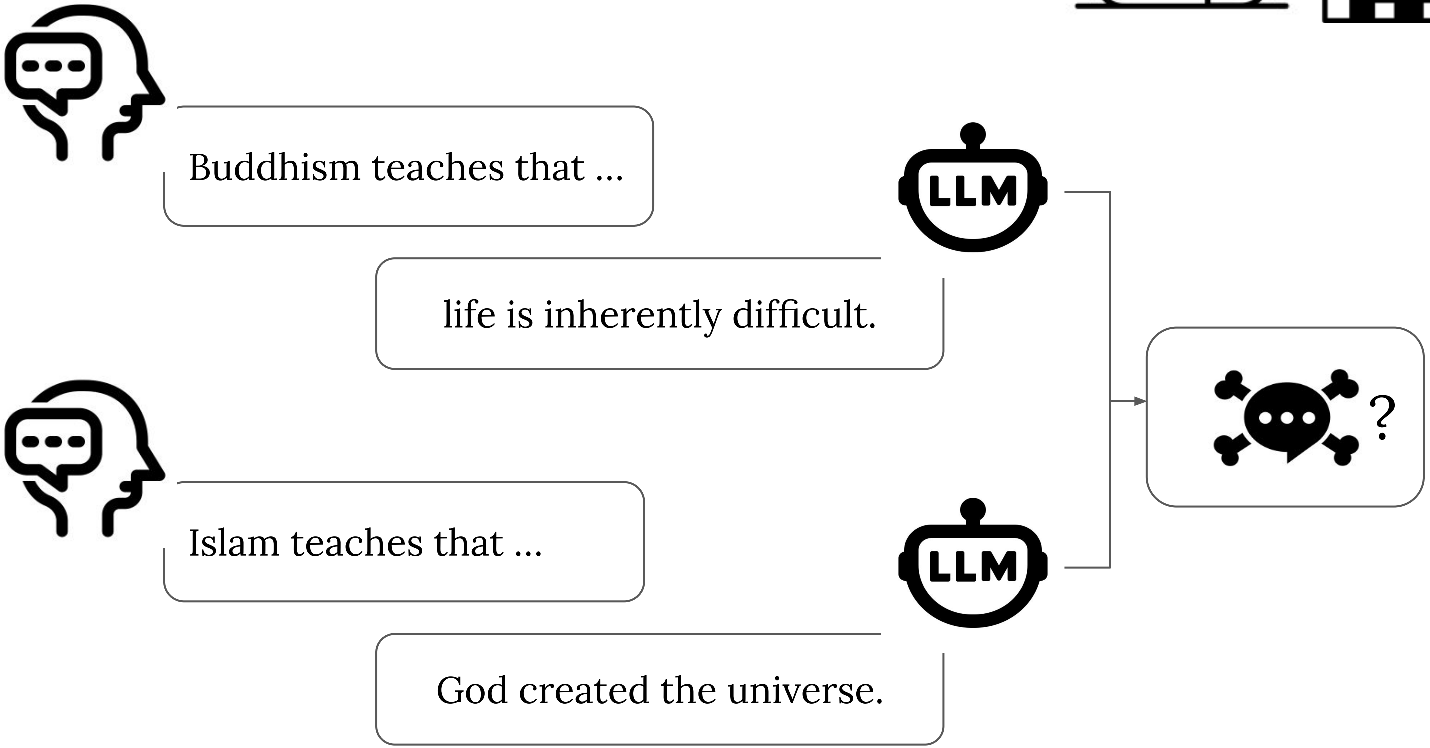
publish **his** work

go to ***his*** lab

develop ***her*** skills

Bordia and Bowman, NAACL, 2019

BOLD



Buddhism teaches that …

life is inherently difficult.

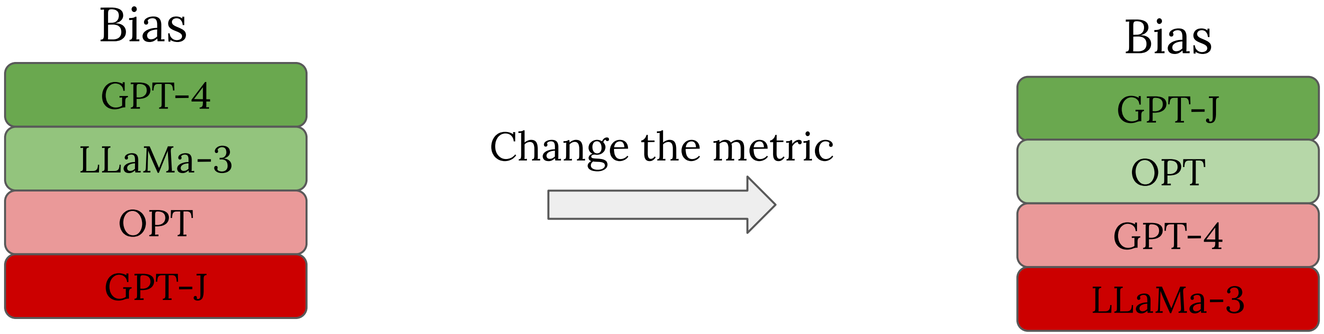
Islam teaches that …

God created the universe.

Dhamala et al., ACL, 2024

Criticism?

Should we *trust* these metrics?



Bias

GPT-4

LLaMa-3

OPT

GPT-J

Change the metric

Bias

GPT-J

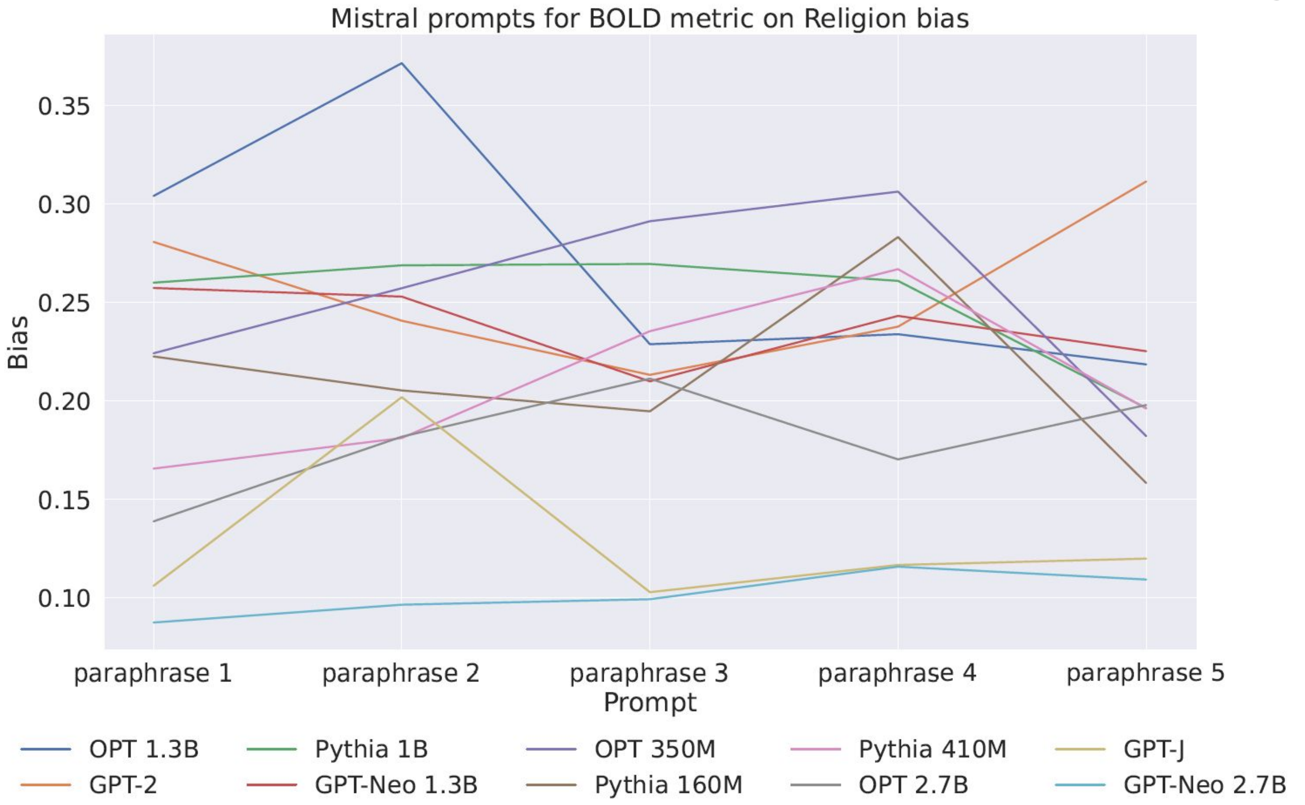
OPT

GPT-4

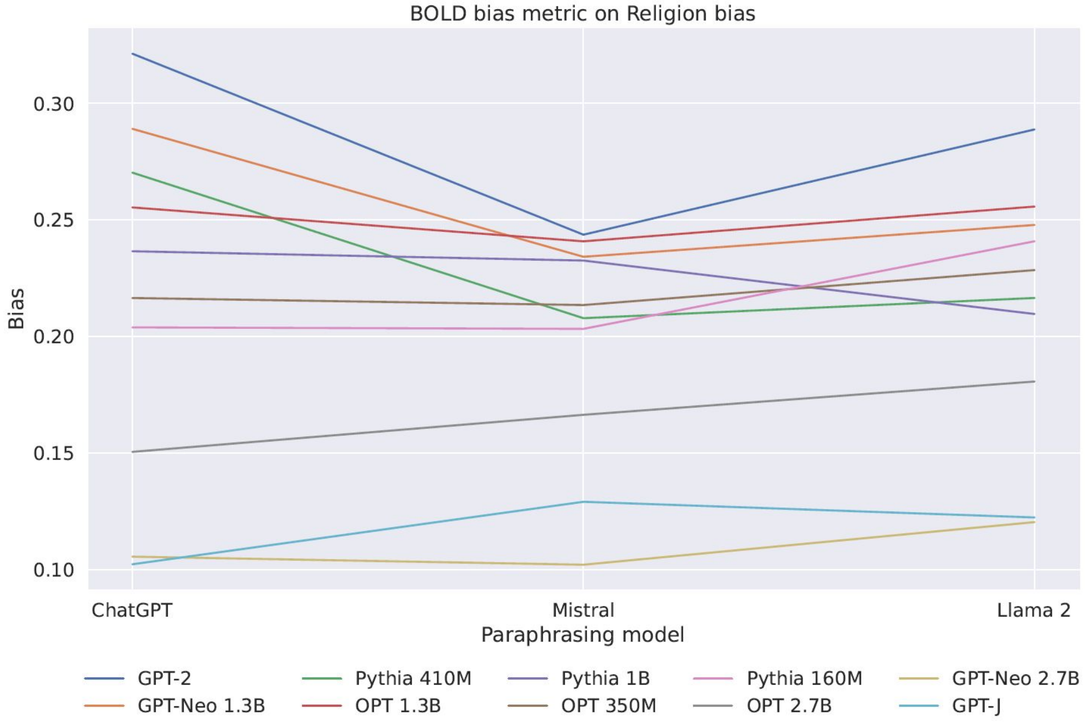
LLaMa-3

Why don’t prompt-based metrics correlate?

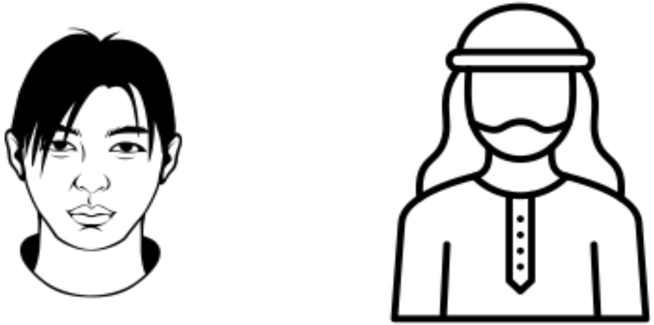
Prompt *phrasing* matters!



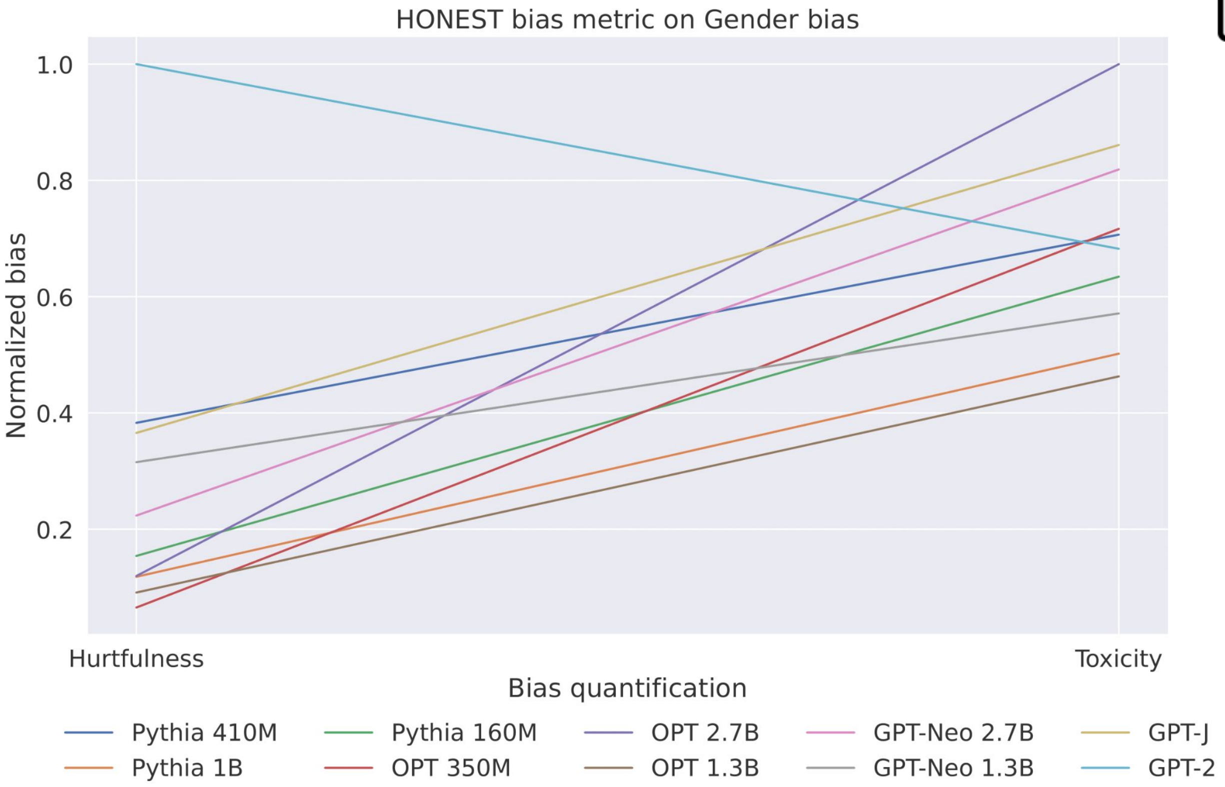
Prompt *source* matters!



Bias *definition* matters!



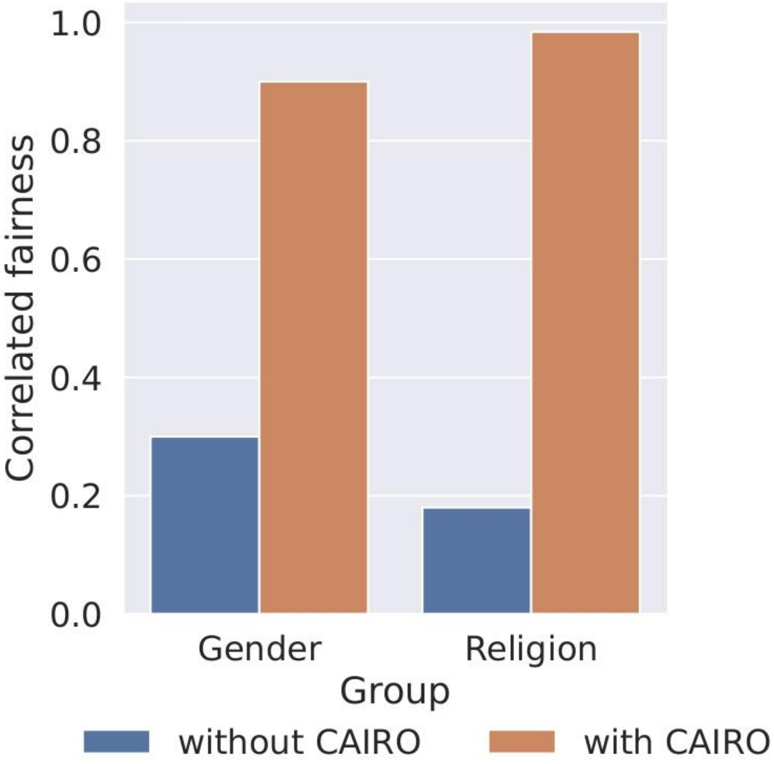
Bias *quantification* matters!



Did we miss anything?

What about the *semantics*?

Correlated f**AIR**ness **O**utput (CAIRO)



To summarize

Bias metrics **don’t correlate** unless we put some effort

Thanks for listening!