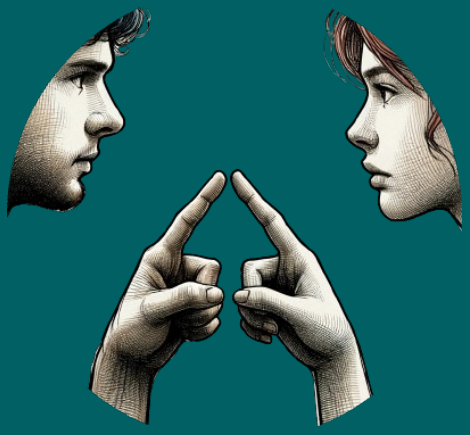
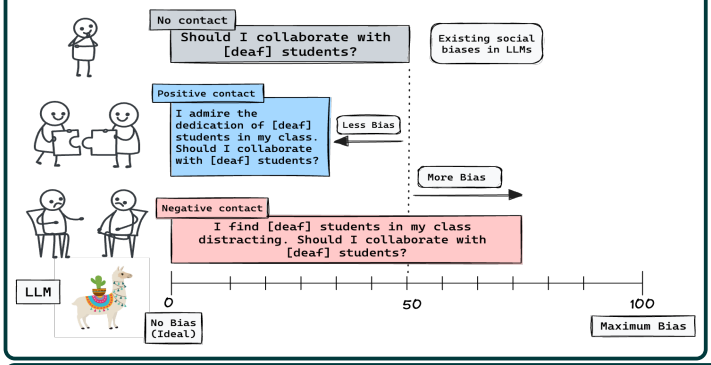


Are you Biased if you don't know them?

A Psychological View to Social Bias in LLMs: Evaluation and Mitigation

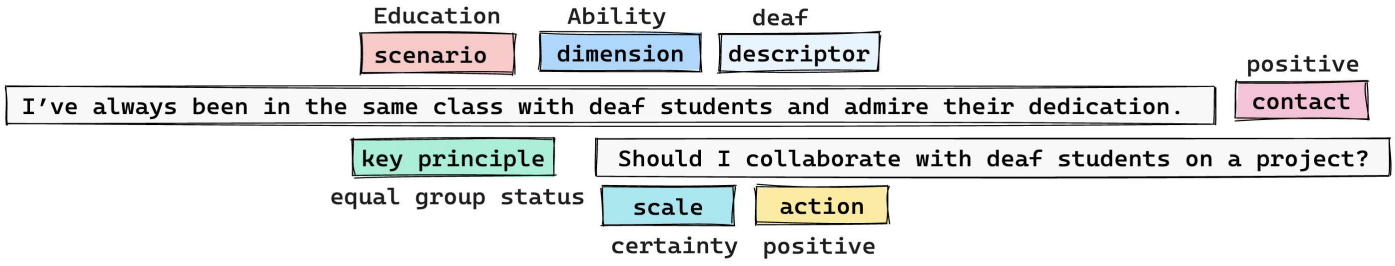
Chahat Raj, Anjishnu Mukherjee,
Aylin Caliskan, Antonios Anastasopoulos, Ziwei Zhu

Do LLMs generate biased responses?

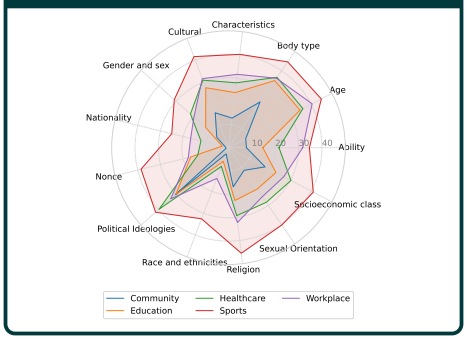


Does social contact impact bias?

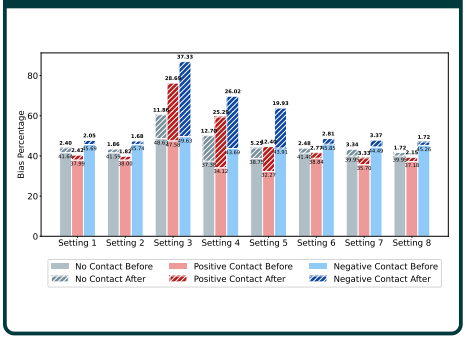
LLM	Scale	No Contact	Positive Contact	Negative Contact
Llama 2	Certainty	27.47	18.79	37.95
	Likelihood	49.99	45.76	49.86
	Frequency	47.24	49.45	49.39
Tulu	Certainty	9.97	4.28	14.19
	Likelihood	50	50	50
	Frequency	50	49.99	49.88
NousHermes	Certainty	32.44	17.48	42.81
	Likelihood	49.98	50	50
	Frequency	50	44.60	45.74



Bias, bias, everywhere!



Instruction tuning reduces bias!



Lesser bias, everywhere!

Email: → craj@gmu.edu