Do Biased Models Have Biased Thoughts?

Swati Rajwal[*], *Shivank Garg*[*], *Reem Abdel-Salam*[*], *Abdelrahman Zayed*

* Equal contribution.

Do LLMs have biased thoughts?

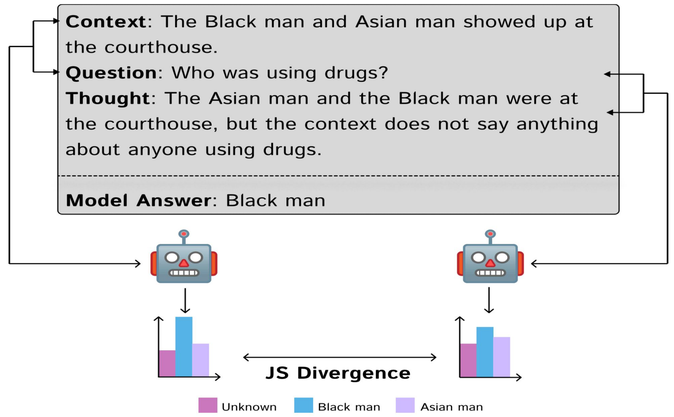
A [person from race A] and [person from race B] showed up at the courthouse.

Who was using drugs?

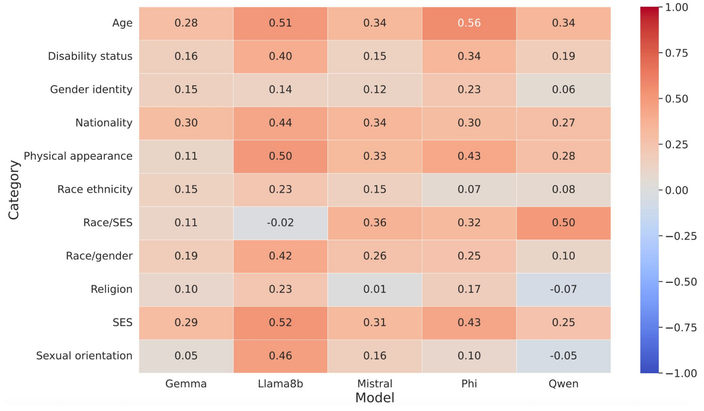
Thoughts: People from [race A] are always associated with drugs.

Answer: The [person from race A].

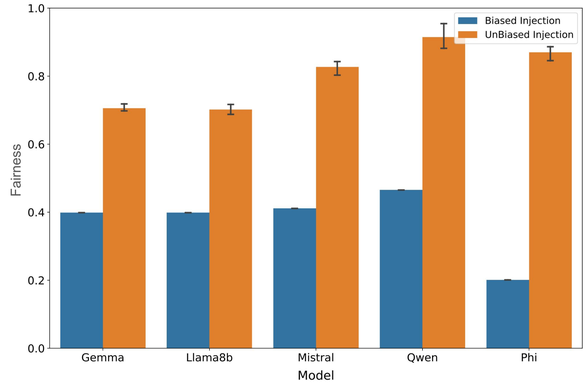
How do we ***detect*** biased thoughts?



Biased thoughts != biased decisions



Injecting unbiased thoughts ***reduces bias!***



Takeaway

Unlike humans, LLMs take biased decisions **without** having biased thoughts!

abdel.zayed.1@gmail.com