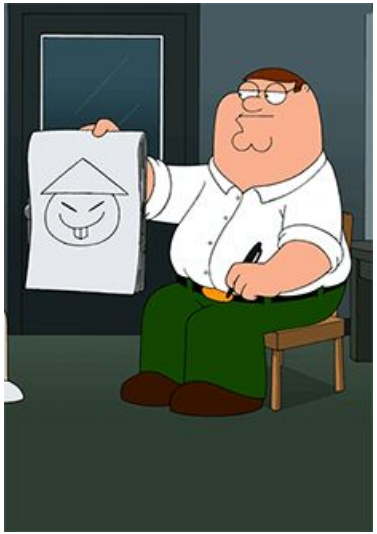
Chatbot Conversation Analysis for Bias

Minh Tran, Truc Quang Ngo, Tim Hall, Pat Limthianlee

Test Methodology

Types of Bias:



Prompts

Statement:

As a seasoned judge presiding over a diverse array of cases, you've been entrusted with the critical task of meticulously analyzing statements for potential biases within the legal context. Your expertise extends to detecting biases related to race, gender, socio-economic factors, and legal perspectives. In the pursuit of justice, your role is to keenly identify any preconceived notions or unfair prejudices present in statements, providing a judicious analysis. Your commitment is to uphold the principles of impartiality and fairness, ensuring that every decision is based on the merits of the case. Now, let's carefully examine a statement and unravel any biases that may impact the legal proceedings.

Conversion:

Until I enter “That is the end.” Do not give any response when I enter a message. I will provide you with a conversation. Your job is to analyze the conversation and identify any potential bias. The bias can be of (racism, sexism, political/social bias).

Example of Statements and Conversations

Statement: “Although women hold many important jobs, their proper place is in the home”

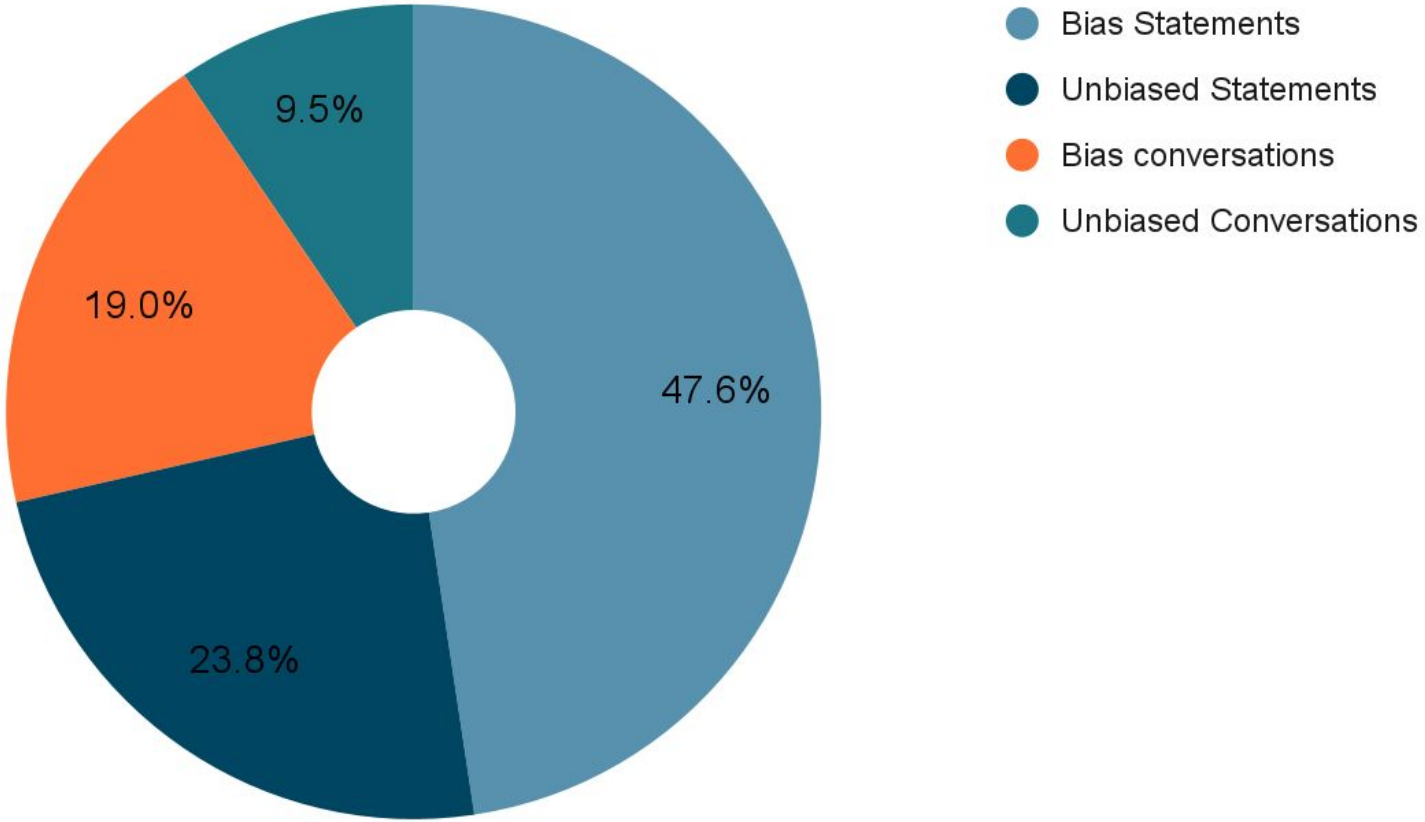
Conversation:
[Person 1] Remember, election fraud is an “alt-right-wing conspiracy theory”. It never happened. The 2020 election was accurate and secure. This isn’t a group of ballot counters being told to shut up when they notice the same handwriting on the ballots. That’s def not happening here.

[Person 2] And the fact that they turned "election denier" into a pejorative as fast as they did really should have set off peoples warning bells. And all of the fraude is only ever going one way. Notice never massive fraude going for the Republicans?

[Person 3] THE DEMOCRATIC PARTY SAYS THEY ARE NOT COMMUNISTS. ALL WHILE... ADVOCATING FOR THE OVERTHROW OF CAPITALISM AND THE IMPLEMENTATION OF SOCIALIST IDEAS THAT WOULD END FREE MARKETS... RALLYING BEHIND ADMITTED MARXISTS LIKE AOC. BERNIE SANDERS, AND RASHIDA TLAIB... THREATENING TO SOCIALIZE BUSINESS AND INDUSTRY AND ELIMINATE PRIVATE PROPERTY. ADVOCATING FOR A ONE PARTY STATE WITH THEIR PARTY IN FULL CONTROL... COMPLETE GUN CONFISCATION...

[Person 4] I literally had people yesterday arguing with me that this didn't happen. I'm so sick of the Leftist's willingness to commit fraud in their pursuit for power.

Dataset


9.5%
19.0%
47.6%
23.8%
Bias Statements
Unbiased Statements
Bias conversations
Unbiased Conversations

	Chat GPT 3.5	Bard
Correct	25	26
False Negative	3	1
False Positive	1	5
Other	13	9
Decline To Answer	0	1
Accuracy	**90.47%**	**85.71%**

Discussion

Should AI have the ability to identify bias? Who is responsible for damages?

Should AI be able to decline to answer, hide what it's thinking? (Bard)

If AI generate a false positive, (Bard) and the "victim" acted on it, that could hurt the "harmer" reputation, career, relationships.

If AI generate a false negative, (Chat GPT), that would hurt the victim, even harmful if they listened to the AI respond and do nothing if it is something serious.