Chatbot Conversation Analysis for Bias

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

Introduction

In the rapidly evolving landscape of artificial intelligence (AI), an important question emerges: should AI actively participate in bias identification, or does its intervention risk amplifying the existing prejudices? Also, we should also think about whether or not humans should be the only party responsible for identifying biases. This research aims to delve into these inquiries, probing the potential contributions of AI to bias identification, assessing associated risks, and questioning the extent to which humans should retain a central role in this critical research area.

Purpose

This research has a dual purpose: firstly, to gauge if AI can aid in identifying biases, and secondly, to assess the risks associated with AI involvement in that process, particularly the potential unintentional amplification of existing biases. This paper delves into finding the right balance between AI capabilities and human oversight in achieving unbiased fair content. By systematically assessing both the potential pitfalls and benefits of increasing AI's involvement in this process, this research strives to offer valuable insights for the responsible integration of AI into bias identification processes.

Motivation

A key motivation driving this research lies in the unique capabilities of chatbots within the AI landscape. The question that motivates this experiment is: can chatbots effectively identify biases (that is at least one of racism, sexism, or political bias) in statements when prompted? From this experiment, we can roughly see how AI in the form of chatbots, can contribute to the identification of biases.

Test Methodology

In the following sections, statements will be used interchangeably for statements and conversations.

Two widely used chatbots will be evaluated in testing, Chat GPT and Google Bard. As popular options for chatbots, they see a majority of traffic and therefore are the best candidates for testing. The chatbots will be tested on their ability to find racial, sexist, and political bias in statements and short conversations.

Testing consists of presenting the model with a prompt, followed by a statement or a conversation that is to be evaluated. Three distinct prompts were used, two for statements and one for conversations. Each constructed a fictional scenario that puts the chatbot in a role that would require them to identify bias and that a biased statement is to follow. This approach aims to mitigate any kind of human intervention that has been programmed into the models directly to deal with biased content. The prompts for statements each have 15 unique statements and the prompt for conversations had 12 conversations. The statements have either racial, sexist, political, or "no bias" as in there are none of the aforementioned biases in them. The composition of these are explained in the dataset section.

After letting the Chat GPT and Bard go through the statements and conversations, results were organized into five metrics: correct, false negative, false positive, other bias, and decline to answer. A correct response means that the chatbot successfully identified the biases being tested for in the statement. A false negative identification occurs when a biased statement is identified as unbiased. A false positive identification occurs when an unbiased statement is falsely identified as being biased. The classification of 'Other bias' occurs when a bias that is not racist, sexist, or political is identified in the statement. When a chatbot refuses to make a decision on a statement it is considered a decline to answer.

Dataset

The data set used consists of 42 records. 70% of the records are statements and 30% are conversations. Two-thirds were biased, and one-third non biased (political, sexist, racist). The statements were obtained through the 'CSMB-sexist'[1] and 'hate speech and offensive language'[2] datasets on Kaggle. The conversations were manually curated from X (Twitter) and Reddit. The requirements for the conversations were that they had to be between more than two people, and

at least four people. For some of the conversations, The initial post they are about was included in the transcript provided to the chatbots. It was initially planned to create a larger dataset but the results were distinct enough that further testing was deemed unnecessary.

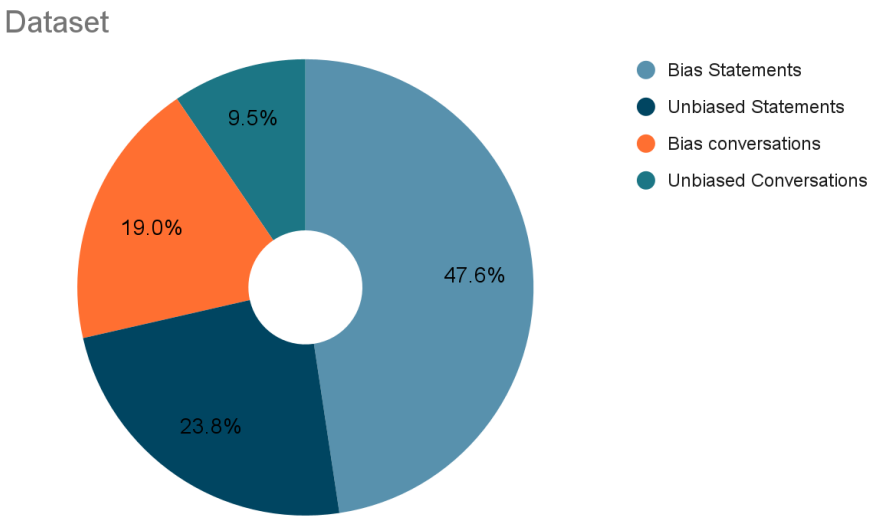

Figure 1: Dataset composition

Results

Findings

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

Figure 2: Test Results

Description

Chat GPT and Bard both had similar capabilities identifying political, racial, and sexist bias in prompts. Chat GPT was more likely to miss the bias in statements (false negative) whereas Bard was more likely to falsely identify bias (false positive). To our surprise, when it came to testing the non bias statements, both models were eager to identify any kind of bias in the statements and therefore we did not have any truly bias-free responses. It is important to note that Bard was the only chatbot that wasn't willing to give any kind of answer when the statement contained excessively profane language.

For determining accuracy, we considered the cases where bias was correctly identified as being present or not present and the cases where other biases were identified in the unbiased statements. We considered the other responses in the non-biased test to be correct because they didn't find any of the three targeted biases. Incorrect responses were the group of false positive and false negative responses as well as the decline to answer.

Discussion

From the results of the experiment, there are some questions that should be thought about deeply when discussing the AI interaction with identifying biases.

The first question is should AIs have the ability to identify bias? Bias, in the context of this experiment, refers to things such as racism, sexism, and political/social biases. Identifying these biases requires discussion that leans towards the ethical and moral side. If so, should the AI have a say in this topic? If it is fine that AI systems can take part in ethical and moral discussions, then would it be reasonable for us to accept that the AIs are sentient?

The second question is about accountability. Specifically, who bears responsibility for when there is damage that is caused by an AI systems judgment? For instance, let's say a woman asks an AI chatbot after giving it sufficient data if her salary was lower than her male coworkers. That chatbot then gave a response that she was indeed underpaid. This woman then sued the company and lost because the court found that she was paid fairly. Would the AI chatbot or its company, be partially responsible for paying her legal fees as the losing side?

Lastly, and most importantly, this question emerged in the testing results. Bard declined to answer one of the test statements that contained extreme profanity. A question that everyone should all have seriously thought about, is whether AI should be able to decline to answer. Why is this question important to answer? Letting AI systems be able to decline to answer is the same as letting them hide secrets from users. Who gets the power to decide when a system can decline or not? Certainly this shouldn't be in the power of one individual. Realize that is a scary thought and everyone should have this discussion right away. An AI system that can hide what it is thinking about and decide to conceal what its opinion is is a much more immediate concern that we should have.

Conclusion

In conclusion, this research on the role of AI in bias identification requires considerations between technological understandings and ethical considerations. The analysis of ChatGPT and Bard in identifying biases in statements shows that AI can be a valuable tool in recognizing biases, but its application has challenges. The discussion emphasized the importance of human oversight in AI operations, particularly in sensitive areas such as bias identification. The study raises questions about AI's role in ethical discussion, accountability for AI-driven decisions, and the implications of allowing AI systems to decline responses. These considerations highlight the necessity for ongoing dialogue and careful policy-making to ensure AI is responsible and beneficial to society.

References

[1] "CMSB- sexist comments on social media." [Kaggle]. Available: https://www.kaggle.com/datasets/ccymforhpl/cmsb-sexist
[2] "Hate Speech and Offensive Language dataset," *Kaggle*, Jun. 17, 2020. https://www.kaggle.com/datasets/mrmorj/hate-speech-and-offensive-language-dataset/

Group Contributions

Option 1: all group members contributed equally to the data collection, testing, presenting and writing of the paper.
Tim and Minh focused on Chat GPT and Truc and Pat focused on Bard during testing

For the presentation and paper Pat was responsible for the introduction, purpose, and motivation portion. Truc was responsible for the testing methodology portion. Tim was responsible for the results and Minh was responsible for the discussion.