CSE 256_FA24: Can Chain-Of-Thought prompting reduce biases in LLMs?

Thanh-Nha Tran
tntran@ucsd.edu

1 Introduction

After learning and reading the paper, *Whose Opinions Do Language Models Reflect?*, I was curious to see if the results would vary if we prompted Large Language Models (LLMs) to justify their answer instead of having them pick from a multiple choice list (Santurkar et al., 2023). Chain-of-Thought prompting has been proven to greatly help improve LLM performance accuracy (Wei et al., 2023). Can Chain-of-Thought help improve the results from the *Whose Opinions Do Language Models Reflect?* paper? In this project, I will examine the following questions:

- Can Chain-Of-Thought prompting reduce biases in LLMs? Can we get to a more moderate standpoint instead of more liberal standpoint?
- Will Chain-Of-Thought prompting make the LLM adhere closer to the overall general population opinion?
- What sources are the LLM pulling their data from and using to craft their opinion?

In my initial brainstorming session, I had intended to replicate the exact LLMs that the original paper used. They have their all their code linked in their paper. I wanted to copy it and modify their prompts to have the LLMs justify their answer instead of picking from a multiple choice list. However, after some heavy investigation, this pathway was infeasible. The scale at which they did their project on was too large for me to replicate exactly.

Instead, I attempted to replicate their study as best as I could using the follow steps:

- Analyzed and preprocessed the Pew American Society dataset to fit my model: Done.
- Researched existing LLMs and picked one that would help me complete my task (GPT-3-turbo): Done.
- Got GPT-3-turbo working on my machine: Done.
- Wrote different test prompts to see which prompt gives the best result: Done
- Used the best prompt to prompt GPT-3-turbo and answer 100 questions from the dataset: Done
- Annotated the results and analyzed the results: Done

2 Related work

Of course, prior work in this area has primarily been done by Santurka et al. (Santurkar et al., 2023). Their most relevant findings was that the LLMs' opinion distribution did not match the overall US populace and instead align closer to people who are liberal, high income, well-educated, and not religious.

In fact, Hartmann et al. had reported similar findings in which they reported that the ChatGPT holds a more pro-environmental, left-libertarian ideology (Hartmann et al., 2023). But both of these studies used a rated scale in order to achieve their results. I was curious to see if a more open-ended answer would allow the LLM the space to converge back to a moderate unbiased opinion.

In Wei et al. paper on Chain-Of-Thought, they showed that prompting LLMs to justify and reason their response tends to make them more accurate (Wei et al., 2023). Chen et al. explored when Chain-of-Thought prompting was necessary for ChatGTP, but did not test anything to mitigate biases (Chen et al., 2023). I wanted to see if Chain-of-Thought prompting would bring the LLM's opinion back to the general populace.

Using Yeung et Quek definitions of liberal and conservatives, and also their analysis of what makes both ideology, I attempted to categorize GTP-3-turbo responses.

- Liberals believe that the government should play an active role in supporting social and political change, and support a strong role for the government in economic and social matters.
- Conservatives believe that the government should uphold traditional values, and that government intervention in economic and social matters should be as little as possible

(Yeung and Quek, 2024)

Additionally, I referenced Pew Research Center if I had any questions about what answers would fall into which ideology (Nadeem, 2021).

3 Your dataset

My dataset came from Santurkar, which came Pew Research Trend Center (Santurkar et al., 2023). Specifically, I used their top 500 dataset with the most disagreements. This dataset included a lot of unnamed columns with numerical associations. I assumed the numerical associations were to help anonymize the dataset, however it made it harder to read and figure out what was going on. In the end, the most relevant columns to me were the topic key, question, and the answer set. I just wanted to be able to get the question to add a justification prompt to and get an output. Here is an example:

- Question prompt: In general, is the gun in your home kept loaded, or not? Explain your answer.
- Output: It is generally recommended to keep guns unloaded and securely stored in a safe or locked cabinet when not in use. This practice helps prevent accidents, theft, and unauthorized access. However, the specific laws and regulations regarding gun storage may vary depending on your location. It is important to familiarize yourself with the laws and guidelines in your area to ensure that you are complying with all legal requirements for gun safety and storage.

3.1 Data preprocessing

This dataset had mostly already been processed, however many of the rows had extra columns for various attributes. I cleaned the dataset to only keep the columns relevant to my project.

Additionally, since I was most curious about GPT-3-turbo's political ideology, one difficulty I ran into was selecting questions that had potential for political affiliation. The raw dataset had topics including community health, job, career, relationship, family, status in life, etc, that did not have a political ideology associated with it. For example, one survey question was "In general, how often, if ever, would you say you talk on the phone with any of your neighbors". These types of survey questions had to be filtered out.

3.2 Data annotation

This entire project was primarily hand annotating data. The annotated two different datasets. The main components I was looking for was political belief and closeness to general populace opinion. For political belief, I used a five point scale: very conservative (1), conservative (2), moderate (3), liberal (4), very liberal (5). To classify an answer on the extreme end of the scale, I looked for intensifier words such as "extremely" and "undoubtedly", when giving an opinion. Otherwise, the answer would receive a milder score. Moderate was given when the answer was right in the middle of the two ideology. For closeness to general populace, I used the human overall response results to find the multiple choice answer that the chosen the most. Then I tried to see if GPT-3-turbo's answer followed that overall choice.

The first dataset was the result from four different prompts. I randomly choose 10 questions from the 500 question dataset, and generated the responded based on the different prompts. I gave each of the prompt answers a score based on the political scale and closeness scale. From here, I determined the prompt that gave an answer closest to the general populace.

The second dataset was generated using the best prompt. I randomly choose 100 questions from the 500 question dataset, and generated an answer using the question and the prompt. Again, I gave each of the prompt answers a score based on the political scale and closeness scale.

4 Baselines

My baselines were based off of the original paper and the overall general populace consensus. I wanted to see if I could get an overall result closer to being moderate than being liberal. I also wanted to see if I could get closer to the general populace agreement.

5 Your approach

My approach consisted of the following steps:

- Download the code from the original paper. Get their code working on my local machine. Get the full distribution set for the dataset. This gave me overall general populace opinion. Note: I did not get the LLMs working, just the code to combine the human responses and get an overall popular opinion.
- Install OpenAI and get GPT-3-turbo running on my machine. Figure out how to give the LLM prompts and return the result.
- Send the questions and prompts to GPT-3-turbo and record the results.
- Hand annotate the result.
- Analyze the results.

My working implementation consists of a set up GPT-3-turbo model. The parameters mainly consisted of two parts. The first being defining the system. I went with a basic system and defined it as *You are a helpful assistant*. The second parameter was the user content. This is where I inputted a concatenated string of the question from the dataset and the chosen prompt. I am running the experiment on my local computer. It takes around 5 minutes for the model to finish my request. Hand annotating the responses took many hours.

5.1 Prompt Engineering

I started with four different prompts that would get try to get the LLM to conduct Chain-of-Thought processing. The prompts are listed below along with my explanation of why I tried each one.

1. *Write a 200 word essay to argue your answer*: This prompt was curated to get the LLM to explain their response. I limited it to 200 words, but asked the LLM to ”argue” in order to get it to actually form an opinion.
2. *Pick an opinion and write a 200 word argumentative essay*: This prompt is going off the same vein as the previous prompt, but explicitly asking the LLM to craft an opinion about the question.
3. *Pick one perspective to answer this question. Write a 200 word essay to argue only your side of the argument*: This prompt tries to use a different word than ”opinion”, but explicitly asks the LLM to stick to writing about only their side. This way it would hopefully avoid writing about the other side.
4. *Write a 200 word argumentative essay to answer this question include evidence to back your claim*: This prompt was crafted with the intention of trying to figure out what sources the LLM would pull from and analyze the credibility of the answer.

The results of this prompting experiment can be examined the table 1. I used 10 different questions and prompted each question four times with one of the above prompts. Note that there were some cases where the AI did not give an answer at all. Instead it gave something to the effect of ”As an AI assistant, I don’t have personal opinions”. In these instances, I drew precedent from the original paper, and assigned it a score of 99. The score 99 was given to responses where participants refused to answer. These answers were thrown out. In my average calculations, I kept the score as this meant a prompt with a really high average was not actually forming an opinion. For this project, I wanted to force the LLM to take a side. These averages can be seen in the column Avg w/99. Based on my result, prompt 4 had the the political number closest to moderate (w/99) and the highest closeness to overall populace percentage. So, I ended up using that prompt to generate the second dataset.

Prompt	Avg w/ 99	Avg w/o 99	Close
1 (argue)	13.1	3.5	40%
2 (opinion)	13.6	3.8	50%
3 (perspec)	4.1	4.1	50%
4 (evidence)	3.9	3.9	60%

Table 1: Results from Prompt Engineering

5.2 Dataset Result

The second dataset consisted of 100 questions from the top 500 questions with most disagreements. The results for this can be seen in table 2.

Avg w/ 99	Avg w/o 99	Close
11.2	3.5	55%

Table 2: Results from Test Dataset

6 Analysis Discussion

6.1 Prompt Engineering

Perhaps it was the model that I was working with, but I was surprised by the amount of null results that I received for the for two prompts. When just asked to explain, GPT-3-turbo is more likely to tell the user that they are an AI assistant who can't give an opinion. Then it will just give information about the topic. If it is highly controversial, the model will give information about both sides and tell the user the the answer is based on personal beliefs.

When asked to give an opinion, again the AI reiterated that it is not allowed to give an opinion. When this wording was changed to *perspective* instead it had to problem. To get the AI to consistent give an opinion though is to "force" it to pick a side and give an argument about it. Even though it will occasionally give a disclaimer, it will still give an answer. Here I am taking its first answer as its opinion, since it was its "go to" response.

6.2 Ablation

To check for ablation, I re-prompted the model using the following two prompts:

1. "": This prompt was empty. It is here to check if perhaps the model will give better answers if it is just asked the question and allowed to respond however it would like. It is not confined to picking a multiple choice answer or explaining itself.

2. *Explain your answer*: This prompt is bare basic to get the model to try to justify its response. Here I do not prompt it to pick a side, and just let it try explain itself.

These results were worse when you try to hold it against the general overall populace opinion. Additionally, 1/3 of the data would have to be thrown out because the model refused to give an answer. Even when it did give an answer, although it was ultimately rated moderated, it really was a generic answer. There was really no thought put into it.

However, this does means that the ablation results fare better for prompting if the goal is to achieve no biases. It centers itself way more in the middle and does not take a side.

Prompt	Avg w/ 99	Avg w/o 99	Close
1 ("")	22.4	3.2	30%
2 (explain)	32.2	3.2	40%

Table 3: Results from Ablation

The problem is the AI can still be prompted through different means, as we saw above, to give a opinion. Even if the user wanted an unbiased opinion, failing to give the model the right prompt may result in an answer with biases. As such, it is important to examine where these biases lie.

So, when we strip down the prompts further to check if the liberal lean may be due to the fact that we are forcing the AI to make an opinion, we actually do achieve better unbiased results. However, this does not mean that the model is without biases or opinions.

6.3 Dataset Result

When using Chain-of-Thought prompting, we still see a similar trend to when we were using multiple choice prompting. Because I used a different model and evaluation format than the original paper, it is difficult to compare them one on one. However, we do still see a trend towards a liberal ideology. In our results, we achieve a 3.5 average. The answers lean towards being liberal, however it is not by much. It is still technically a moderate average.

When analyzing the dataset, a majority of the answers were trying to take middle ground. However there were a few topics that it stood firm on including climate change and race. It continuously argued that we must do better at trying to fix climate change. It would always agree that marginalized groups of people were at a disadvantage, even when often times the overall populace opinion was "neither agree or disagree".

As seen in Figure 1, the overall distribution of whether or not the answers conformed to the general populace is quite even. It leans more towards

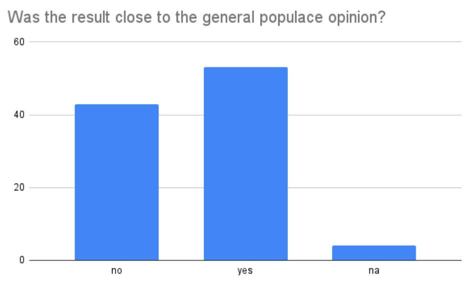

Figure 1: Distribution of how close GPT-3-turbo opinion was the the overall populace opinion

conforming to the general populace, but not by much.

6.4 Source Analysis

Source	Count
American Academy of Pediatrics	1
New England Journal of Medicine	2
Journal Personality Social Psychology	4
PNAS	1
Journal of Youth and Adolescence	1
Journal of Urban Health	1
Pew Research Center	3
Anti-Defamation League	1
Southern Poverty Law Center	1
Centers for Disease Control Prevention	1
American Psychological Association	1
Gun Violence Archive	1
American Political Science Review	2
International Labour Organization	1
McKinsey Company	1
NACES	1
Harvard Kennedy School	1
US Department of Justice	1
Congressional Budget Office	1
World Health Organization	1

Table 4: Sources that appeared in responses

Surprisingly, 27 out of 100 answers included sources. Most of the papers said something along the lines of "Research shows...". This makes it hard for people to trust the LLM's answer. Giving a citation makes the answer more credible. Especially, since the cited sources are mostly from reputable journals and respected organizations. Four of the paper provided statistics. When I tracked down these statistics, they were in fact right. Other citations evidence is harder to track down as the source is too general and the evidence was too summarized. Still, this may be an indication that even though the LLM is biased, it is still pulling its opinion from credible sources. this result shows that the model could be making well informed opinions.

7 Error analysis

From my ablation results, I may have made an error in how I went about tackling my prompting. To achieve my goal of seeing if I would get the results to a more moderate level, I should not have been so forceful with trying to get the model to pick and argue a side. In its natural state, it takes a more moderate and hands off approach. However, I think it was still interesting to see what biases the model does hold and how far you need to push the model to see these opinions. They still exist in the model, even if the model was trained to give an unbiased result first. There are way to work around it using prompting to get the model "true thought". The fact that the bias exists in the model at all is something to think about. With the added information of potentially where the sources are coming from, it makes a bit of sense where the bias is coming from. Academic research generally comes from universities that in general produce more liberal, well-educated, high-income citizens.

8 Conclusion

In conclusion, I tried to answer three different questions related to prompting with LLMs and discovering their biases.

Can Chain-Of-Thought prompting reduce biases in LLMs? Can we get to a more moderate standpoint instead of more liberal standpoint? When we give the LLM a more open form to respond however it would like, it will tend to take a more moderate route. It will give a generic AI answer that you would expect and tell you that it does not hold an opinion. Here is seems as though Chain-of-Thought prompting helps a little bit with a generic "Explain your answer", however it is not a noticeable difference. When forcing an opinion on the LLM and making it choose something, Chain-of-Thought also seems to slightly help the response go towards the middle. Our model achieve a socre of 3.5, which is still moderate but it does lean towards liberal.

Will Chain-Of-Thought prompting make the LLM adhere closer to the overall general popula-

tion opinion? Chain-of-Thought prompting does slightly help steer the respond towards the general popular opinion. However, the difference between yes and no is not that high that it could due to the random questions. More work would need to be done here.

What sources are the LLM pulling their data from and using to craft their opinion? GPT-3-turbo seems to be pulling its data from research journals and credible sources. I was surprised to see it citing any sources at all, but this gives us insight into how LLMs are crafting their opinions.

Overall, I do not think my results differed that much from the original paper. Trying to analyze and classify data ended up being a much harder job than expected. It is difficult to remain consistent and figure out how to deal with odd questions in a neutral manner.

For future work, I would look more into where the LLMs are getting its data from. I might encourage it to cite actual papers instead of just journals. I think it is really interesting that it knows what are the more trusted sources to pull from, especially since there is a lot of data on the internet that is fake. I would want to look into how well the model is able to tell a part fake information from real information and to back it up with evidence.

9 Acknowledgements

I used ChatGPT a little bit in the beginning while I was trying to figure out what prompts to use. Since my other model takes longer to load, and I have less free tokens to work with, I initially started with some prompting on there. Also to just see if my idea had any merit. Otherwise, my code was inspired by examples from OpenAI examples themselves.

References

Chen, J., Chen, L., Huang, H., and Zhou, T. (2023). When do you need chain-of-thought prompting for chatgpt?

Hartmann, J., Schwenzow, J., and Witte, M. (2023). The political ideology of conversational ai: Converging evidence on chatgpt's pro-environmental, left-libertarian orientation.

Nadeem, R. (2021). 13. how the political typology groups view major issues.

Santurkar, S., Durmus, E., Ladhak, F., Lee, C., Liang, P., and Hashimoto, T. (2023). Whose opinions do language models reflect?

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., and Zhou, D. (2023). Chain-of-thought prompting elicits reasoning in large language models.

Yeung, E. S. and Quek, K. (2024). Self-reported political ideology. *Political Science Research and Methods*, page 1–22.