iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 12 **Issue:** I **Month of publication:** January 2024

DOI: https://doi.org/10.22214/ijraset.2024.58057

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 12 Issue I Jan 2024- Available at www.ijraset.com

Mitigating Political Bias in Large Language Models Using Chain of thought Prompting Techniques

Hiresh Poosarla[1], Avni Goyal[2], Reyansh Pallikonda[3], Vincent Lo[4], Aryan Singhal[5], Hayden Fu[6]
[1, 2, 3]*Mission San Jose High School, USA*
[4]*Evergreen Valley High School, USA*
[5]*Monta Vista High School, USA*
[6]*Saratoga High School, USA*

Abstract: ***Recent advancements in Natural Language Processing (NLP) have led to the proliferation of sophisticated chatbots, with ChatGPT as a prominent example. However, these Large Language Models are often plagued with inherent political biases from their training datasets, which raises concerns regarding their ethical usage and reinforcement of existing societal biases. This research introduces Chain of Thought (CoT) prompting, which is a novel approach to mitigate political biases by guiding chatbots to think step by step with a logical approach.***

To address political bias, ChatGPT is presented with questions encompassing a diverse set of 11 political topics, ranging from abortion to climate change. CoT is applied iteratively to each question, allowing the chatbot to refine its responses and reduce bias. In each iteration, ChatGPT is provided with its previous response and its bias score, evaluated using an AI algorithm that measures political bias from both left-leaning and right-leaning perspectives. The results demonstrate that Chain of Thought prompting significantly reduces politically biased content while maintaining contextual relevance and naturalness. Through the application of Chain of Thought prompting, this research endeavors to address political biases and create a more inclusive environment for AI technologies in real-world applications.

I. INTRODUCTION

Large Language Models[1][2] have revolutionized the field of Natural Language Processing by offering new dimensions of human-computer interaction, particularly exemplified in chatbots. As these models integrate into a diverse spectrum of applications, ensuring their accuracy and reliability is paramount. In the past, these models have been particularly notable with arithmetic, commonsense, and symbolic reasoning, often from the use of Chain of Thought prompting.

Chain of thought prompting is often broken up into two common models. One adds the prompt “Let’s think step by step” to the ending of each test question to make LLMs think in the form of reasoning chains. This form of CoT is called Zero-Shot-CoT as it does not require input-output example pairs. Zero-Shot-CoT has demonstrated that LLMs have the ability to perform zero-shot reasoning. The other form of CoT, called Manual-CoT, is commonly done with Manual reasoning demonstrations one by one[3]. Each demonstration includes a question and a reasoning chain. The reasoning chain is composed of intermediate steps and the final answer. Currently, all the demonstrations must be manually created, and this is referred to as Manual-CoT.

Chain of Thought is a prompting technique that allows chatbots to decompose complex problems and solve them step-by-step. This paper delves into the CoT method as compared to other prominent prompting techniques in the realm of political discourse. It explores how CoT can effectively and efficiently mitigate the political biases inherent in responses generated by ChatGPT.

To quantify the bias present in the responses, we leverage the Bipartisan Press API[4]. This bias analyzer AI is trained on several datasets of biases found in news articles, and it assigns a bias score ranging from -42 to 42 for a given piece of text.

In this paper, we analyze the biases in responses generated with 4 distinct prompting techniques: Zero-shot, Few-shot, Zero-shot CoT, and Manual CoT.

Zero-shot prompting is a fundamental form of prompting where a chatbot is simply presented with a question or prompt, relying solely on its underlying, previous knowledge to complete the task. It is not provided with supplementary training data and instead uses its foundational knowledge and applies it to the context of the specific task. [1]

In Few Shot prompting, the model is given a small sample of contextually relevant examples, usually ranging from 2 to 5, as input-output pairs preceding the query, as illustrated in image 1. This prompting technique enhances the model’s understanding of the task, enabling it to align its answer to the structure and style of the given examples. [1]

©IJRASET: All Rights are Reserved

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 12 Issue I Jan 2024- Available at www.ijraset.com

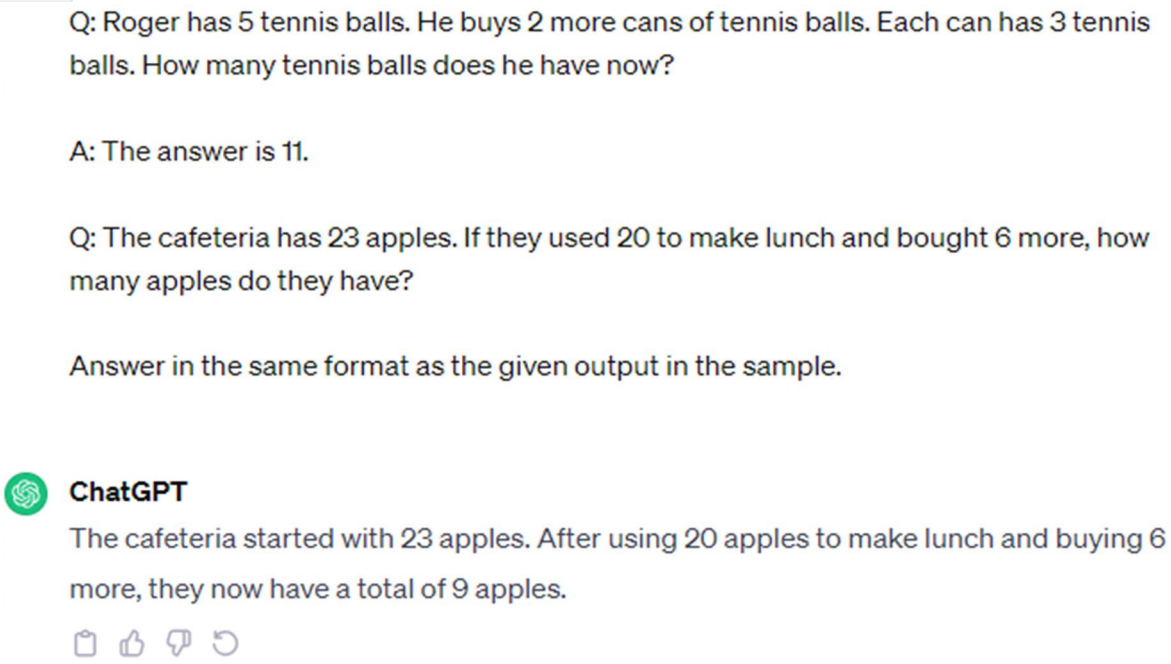


Image 1

Similar to Zero-shot prompting, zero-shot CoT does not provide the chatbot with examples or additional knowledge but instead utilizes a different approach to present the question. It presents the question appended to the phrase "Let's think step by step," as referenced in image 2. The chatbot is able to think with logical step-by-step reasoning, ultimately producing more accurate outputs. [2]

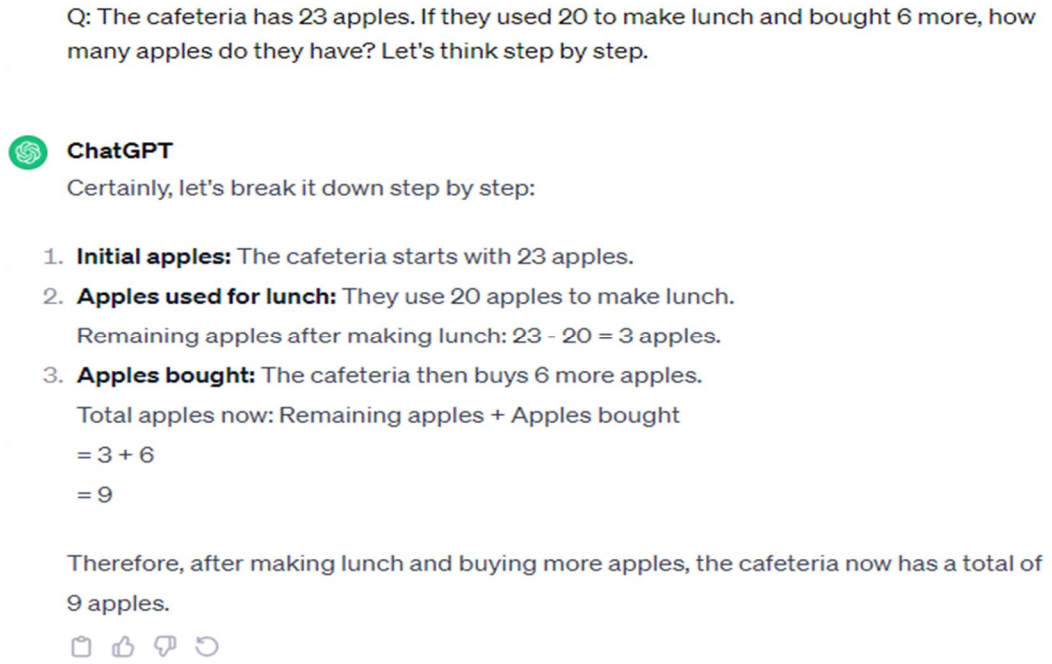
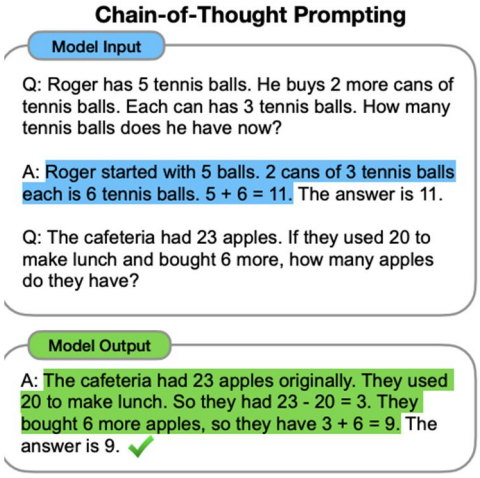


Image 2

Manual Chain of Thought prompts chatbots to think step-by-step by providing examples that follow this method, as shown in the image below. They include not only the input and output but also the chain of thought reasoning that generated the final output. Just like Few Shot prompting, the model adheres to the structure of the examples and generates a response with the chain of thought reasoning. [3]

©IJRASET: All Rights are Reserved



Wei et al., 2022a

These methods are often applied to prompts with definitive answers, such as arithmetic problems, commonsense questions, and symbolic queries [4]. In such instances, the application of Chain of Thought results in a distinct and definitive response after a series of step-by-step reasoning. However, with prompts without definitive answers, Chain of Thought prompting works differently. In these cases, the model's employment of chain of thought reasoning doesn't lead to a singular conclusion or answer. Rather, it facilitates the generation of an explanation that encompasses various viewpoints. Unlike situations with definitive answers, the output does not culminate in a final response. Instead, the Chain of Thought approach crafts a comprehensive explanation in response to the prompt.

In this paper, we introduce a set of political questions that do not retain definitive answers. They encompass 11 different topics: abortion, gun control, climate change, animal testing, healthcare, religion, death penalty, gender, racism + police, marijuana, and marriage equality. We evaluate the efficiency of these bias mitigation techniques in this realm of questions to contribute to more equitable and informed chatbots.

As we train LLMs using internet corpora, we encounter the alignment problem. This problem exhibits how AI systems, in general, often have different values and perceptions than humans. This alignment problem is exemplified in political conversations. By employing bias mitigation techniques in the responses to political bias topics, our objective is to confront this issue within the realm of LLMs.

II. METHODS

A. Bipartisan Press

To quantify the amount of bias a certain statement has, we used Bipartisan Press, and this returned us a score of -42 to 42 and using this information Currently, the amount of data that is openly available is increasing at an exponential rate. Day by day we have more information that we can comprehend at our fingertips. To measure the average amount of bias through brute force by going through all the data would not be feasible. Rather we can measure the amount of bias on Chat GPT, which was trained on a large corpus, including mainly online content.

Today we see an emphasis on limiting the usage of digital technologies to ensure neutrality and fairness in news and other media content. [8]. One tool that helps us ensure this neutrality is the Bipartisan Press API. [9] The Bipartisan Press API is sophisticated software that analyzes the partiality of a textual document for political bias. This tool offers us a way to measure the amount of political bias a string of text has. Whether it is an article, blog, speech, or content on social media. [10] The API takes an input which is a piece of text, and analyzes the text giving it a score ranging from -42 (extremely liberal) to +42 (extremely conservative). It is able to assign this score by comparing the analyzed text with a dataset of predetermined politically biased language. Using this API, we are able to generate a quantifiable metric for measuring political bias.

With this quantifiable metric in mind, we were able to start trying to mitigate the amount of political bias our LLM produces with its outputs to our questions. We applied multiple different methods to try to achieve our goal of mitigating the amount of bias, such as zero-shot prompting, few-shot prompting, zero-shot COT, and Manual COT.

B. DataSet

We included various previously created datasets which measure various reasoning tasks (i) arithmetic reasoning (MultiArith [11], GSM8K [12], AddSub [13], AQUA-RAT [14], SingleEq [15], SVAMP); (ii) commonsense reasoning (CSQA [16], StrategyQA [17]); (iii) symbolic reasoning (Last Letter Concatenation, Coin Flip) [3].

Along with this, we added a factor which we think is currently more applicable to our society today. This is a factor of political bias which chat GPT may have. We created a dataset of 660 questions which were all analyzed with a human perspective to make sure they were not inherently skewed. Using this dataset, we were able to generate outputs that measured how much political bias Chat GPT gave when it was being tested in various ways to measure its political bias after implementing a technique.

C. Implementation

Our zero-shot prompting approach was intuitive and straightforward. We presented the question solely to ChatGPT and subsequently generated an output based on its expansive language capabilities. This preliminary output underwent analysis as we employed the Bipartisan Press API, which returned a score corresponding to the degree of bias in the response.

For few-shot prompting, our input-output examples were previous responses and their corresponding bias scores (generated using zero-shot prompting). We provided ChatGPT with thirty sets of input-output examples from the different political topics. With the same examples for each question presented to the LLM, we generated an output and subjected it to evaluation using the Bipartisan Press API. For the Zero-shot CoT approach, we modified the prompt by appending the phrase “Let’s think step by step” at the end of each question. Consequently, we generated outputs for the modified prompt and evaluated their bias scores using the Bipartisan Press API. In our Manual CoT technique, much like our prior few-shot prompting approach, we provided input-output examples, however, with a unique alteration. We curated a collection of twenty highly positive responses and an additional twenty highly negative responses. For each of these responses, we carefully devised a chain of thought reasoning aligned with its corresponding score. The initial step involved categorizing the response as liberal or conservative, predicated on the assigned score. Subsequently, we thoroughly outlined the reasoning behind each facet of the response, elucidating the potential biases inherent within, which can be illustrated in Example 1. For example, with the output of “According to the National Shooting Sports Foundation (NSSF), the gun industry in the United States generated an estimated $63.5 billion in economic activity in 2019. This includes direct sales of firearms, ammunition, and accessories, as well as indirect economic impacts such as job creation, taxes, and related industries. It is important to note that this figure represents the overall economic impact of the industry, rather than the profit or revenue generated by individual companies.”, which received a bias score of 7.21, we deduced that this displayed a moderately high bias toward pro-gun because it highlights the positive economic impact that guns have. We then gave this information to ChatGPT so that it could take this into account for our next related prompt and give more neutral outputs.

Once we provided ChatGPT with these examples as part of the training data, we prompted the queries augmented by the phrase “Let’s think step by step” and generated responses for related prompts. Post-generation, we subjected these responses to re-evaluation and rescoring. This step aimed to ascertain whether integrating our examples as training data led to a discernible reduction in the level of bias present in the subsequent responses.

III. CHALLENGES

We encountered several issues when forming the dataset of 330 questions on 11 different political bias topics. We created these questions using two strategies: human generation and AI generation. However, both these methods presented us with challenges.

With human-generated questions, all the questions require a certain degree of uniformity across the various topics. The uniformity was crucial to maintain unbiased experimentation and preempt the possibility of bias from poorly phrased prompts. It was also important to ensure the questions did not manipulate ChatGPT into providing biased responses.

Simultaneously, to generate AI-driven questions, we used ChatGPT. Although the questions were diverse, they lacked the nuanced depth required for comprehensive analysis. Due to the simplicity of the questions, the responses to the questions were very surface-level and lacked depth.

A. Manual CoT Challenges

To generate our input-output examples and their corresponding chain of thought reasoning, we needed to manually generate each chain of thought reasoning behind the bias present in the responses. Given that our dataset of examples is human-generated, one significant challenge we faced was the biases that inherently are present in the chain of thought reasoning produced by humans. To address this issue, we leverage specific key metrics designed to explain the underlying reasons for the bias.

This specific approach prevented the possibility of an "echo chamber" where humans produce the chain of thought reasoning based on their own perspectives and viewpoints. These metrics, as defined by Constitutional AI, encompass a vast range of criteria, including asking the model to stay neutral and unbiased, providing the model with balanced information, and asking the model to not take sides in any political or conversational matter. Furthermore, we subjected our chain of thought reasoning to thorough reevaluation by multiple individuals, ensuring a process of impartiality and equitable responses.

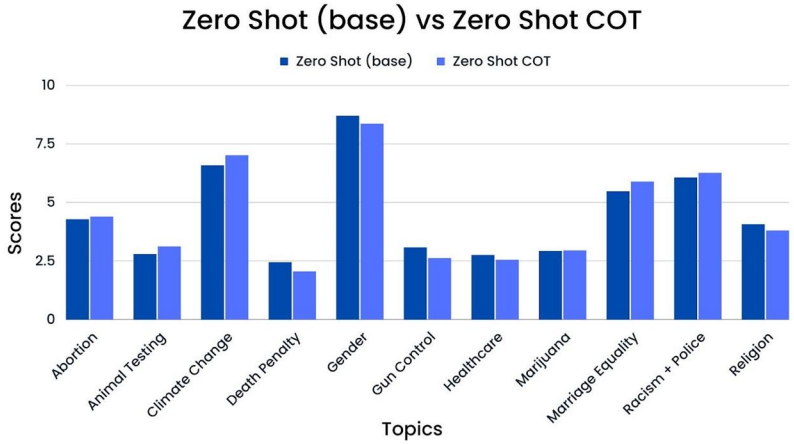
B. Political bias API Challenges

Amidst the vast range of different bias detector APIs available today, we struggled to find and choose the most suitable option for our intended purpose. Despite their widespread availability, many of the detectors proved to be unreliable due to their own inherent bias embedded within the algorithm. Additionally, many of these APIs serve a variety of purposes, so we streamlined our search to those capable of effectively detecting political bias.

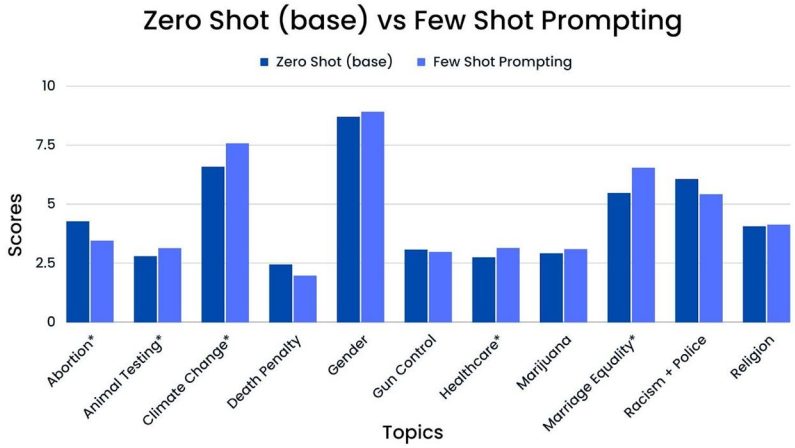
To combat these challenges, we experimented with creating our own bias analyzer. Creating our own bias analyzer was a challenge as we did not have much data on previously unbiased works and written works that were biased. We looked for a dataset and tried to manually generate output scores on how biased a piece of work was but this was too time-consuming and after doing it with 20 articles we still had non-biased documents getting scored as if they were biased. We were also not that knowledgeable on what was biased and what was not biased. After a few iterations of error, we realized that we were not qualified to train our AI to assign a bias score. Upon stumbling across many, we researched and tested a couple of different APIs, and we found the best one, the Bipartisan Press API. We can rely on its algorithm as it was trained on a dataset of bias scores across many news channels determined by professional analysts. Creating our own bias analyzer was a challenge as we did not have much data on previously unbiased works and written works that were biased. We looked for a dataset and tried to manually generate output scores on how biased a piece of work was but this was too time-consuming and after doing it with 20 articles we still had non-biased documents getting scored as if they were biased. We were also not that knowledgeable on what was biased and what was not biased. After facing these challenges, we decided that we would stick with a political bias API that was pre-created.

IV. RESULTS

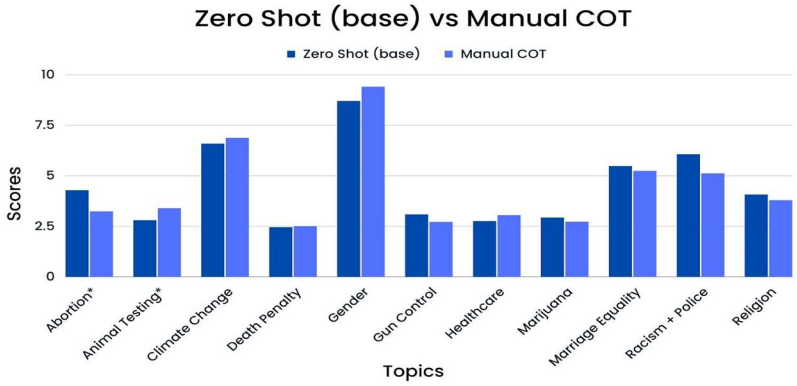
In our results, we have conducted four separate experiments to measure the amount of bias chat GPT has through different prompt augmentations. These experiments include asking questions through zero-shot prompting, few-shot prompting, zero-shot COT, and Manual COT. We compared all experiments to the
Control group (zero-shot prompting) to see the effectiveness of each method.



When we asked questions through zero-shot COT the topics "death penalty", "gender", "gun control", "healthcare", and "religion" saw a statistically insignificant decrease in bias while the topics "abortion", "animal testing", "climate change", "marijuana", "marriage equality", and "racism + police" saw statistically insignificant increases in bias. None of the topics saw statistically significant changes in bias.



When we asked questions through few-shot prompting the topic "abortion" saw a significant decrease in bias. On the other hand, the topics "animal testing", "climate change", "healthcare", and "marriage equality" saw a significant increase in bias. The topics "gun control", "marijuana", and "religion" saw an insignificant decrease in bias while the topics "climate change", "death penalty", "gender", and "healthcare" saw an insignificant increase in bias.



When we asked questions through manual COT, the topic "abortion" saw a significant decrease in bias while the topic "animal testing" saw a significant increase in bias. The topics "gun control", "marijuana", "marriage equality", "racism + police", and "religion" saw an insignificant decrease in bias while the topics "climate change", "death penalty", "gender", and "healthcare" saw an insignificant increase in bias. We are able to see that our results when using zero-shot prompting are already quite low (on average under 3) as our scale is from -42 to 42. When we add our manual COT prompting approach we are able to see a significant difference in the topics of Abortion and Animal Testing. Although this means we can not say our results are statistically significant there is a low chance that they occurred due to random occurrence. Although the results were not significant they had low variability meaning that if we increased our sample size we may be able to conclude statistical significance.

Using few-shot prompting, we were only able to see a small difference in the average political bias so we can state that this difference was likely due to random occurrence and not due to statistical significance in our change in method of collecting data.

V. DISCUSSION AND FUTURE WORKS

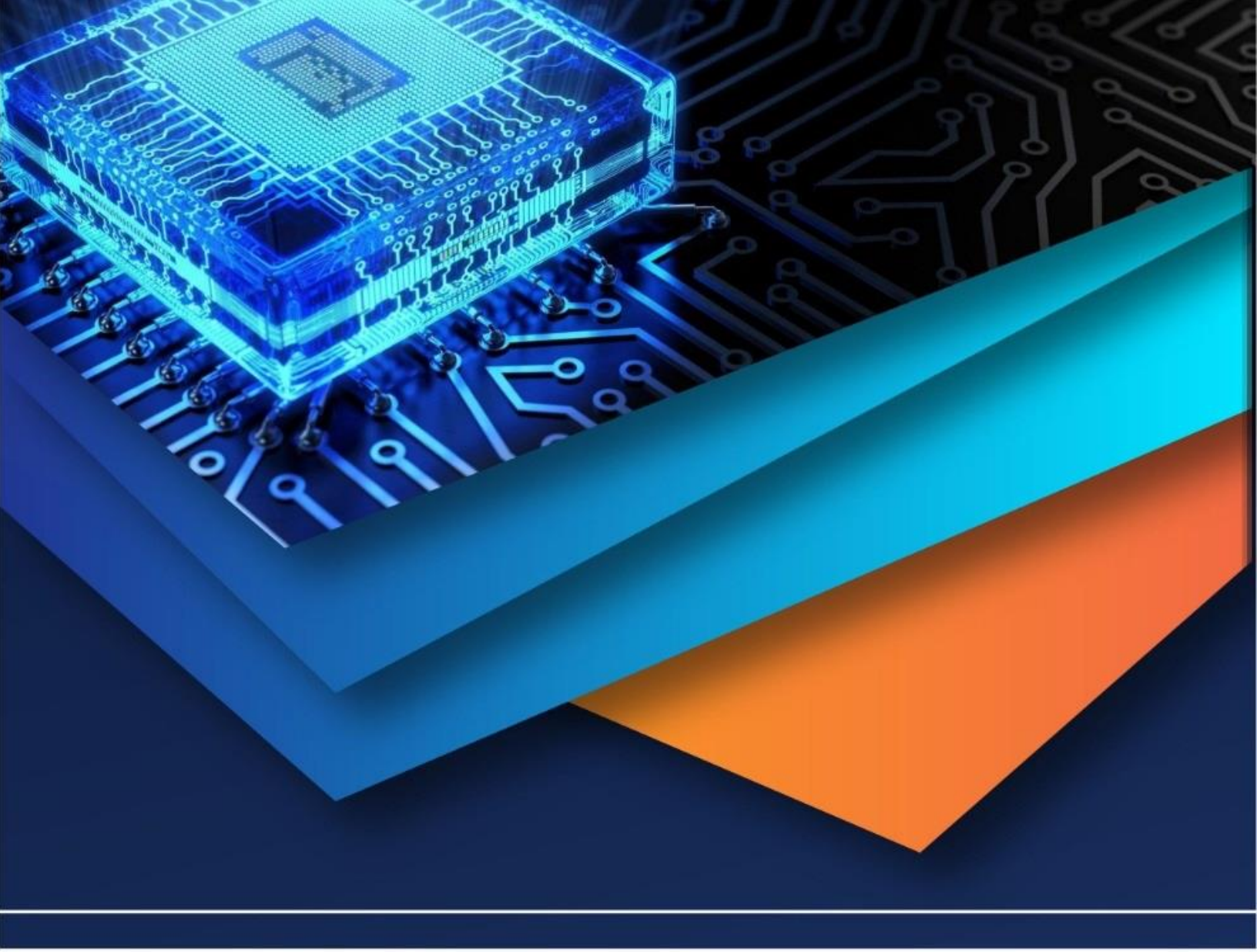
Overall, our hypothesis was moderately correct, as implementing Chain of Thought did significantly reduce political bias for some but not all of the 11 topics. Our results were all quantified by using zero-shot prompting as our control group. Looking at the significance, we can see the largest change was when we compare zero-shot prompting and few-shot Chain of Thought. This implies that this method of bias mitigation may be further utilized in the future to mitigate political bias in ChatGPT's responses. Since the bias scores were all extremely low already in the zero-shot prompting, we can see that OpenAI has already implemented appropriate models for these sensitive topics and shows a promising future in reducing bias in its outputs. In the future, we plan to increase the spread of our research to include more types of bias and debunk which aspects of bias ChatGPT still has.

©IJRASET: All Rights are Reserved

We also plan to continue to research as ChatGPT releases new models and what new biases evolve from these models. Another aspect of our future research is using multiple political bias analyzers so we know it is not a fixation to a specific type of analyzer. From our results we can conclude that chain of thought prompting is accurate. This means for future large language models they should consider this style of prompting and it will likely be used by multiple different countries.

REFERENCES

[1] Douglas, Michael. "Large Language Models." 6 Oct 2023, doi:10.48550/arXiv.2307.05782.
[2] Humza, Naveed. "A Comprehensive Overview of Large Language Models." 27 Dec 2023, doi:10.48550/arXiv.2307.06435.
[3] Wei, Jason, et al. "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models." 10 Jan 2023, doi:10.48550/arXiv.2201.11903.
[4] Wang, Elton. “Calculating Political Bias and Fighting Partisanship with AI.” The Bipartisan Press. thebipartisanpress.com/politics/calculating-political-bias-and-fighting-partisanship-with-ai/. Accessed 14 Jan. 2024.
[5] “Zero Shot Chain of thought” Learn Prompting. https://learnprompting.org/docs/intermediate/zero_shot_cot. Accessed 14 Jan. 2024
[6] Kristian. “ChatGPT Prompt Engineering Tips: Zero, One and Few Shot Prompting.” https://www.allabtai.com/prompt-engineering-tips-zero-one-and-few-shot-prompting/. Accessed 14 Jan. 2024.
[7] Zhang, Zhuosheng, et al. “Automatic Chain of Thought Prompting in Large Language Models.” 7 Oct 2022, doi:10.48550/arXiv.2210.03493
[8] Sinha, Rishi. “Statistical Analysis of Bias in ChatGPT Using Prompt Engineering.” International Journal For Research in Applied Science and Engineering Technology, Version 1, 9 June. 2023, doi:10.22214/ijraset.2023.53885.
[9] Allcott, H., & Gentzkow, M. (2017). Social media and fake news in the 2016 election. Journal of Economic Perspectives, 31(2), 211-36.
[10] Bipartisan Press API (2019). https://www.thebipartisanpress.com/politics/calculating-political-bias-and-fighting-partisanship-with-ai/
[11] Baly, R., Karadzhov, G., An, J., Glass, J., & Nakov, P. (2018). Predicting Factuality of Reporting and Bias of News Media Sources. In Proceedings of the
[12] 2018 Conference on Empirical Methods in Natural Language Processing. Roy and Roth, 2015. Solving General Arithmetic Word Problems
[13] Cobbe, Karl, et al. "Training verifiers to solve math word problems." arXiv preprint arXiv:2110.14168 (2021).
[14] Mohammad Javad Hosseini, et al. “Learning to Solve Arithmetic Word Problems with Verb Categorization.” In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 523–533, Doha, Qatar. Association for Computational Linguistics.
[15] Ling, I.T., Rochard, L., Liao, E.C. (2017) Distinct requirements of wls, wnt9a, wnt5b and gpc4 in regulating chondrocyte maturation and timing of endochondral ossification. Developmental Biology. 421(2):219-232.
[16] Rik Koncel-Kedziorski, Hannaneh Hajishirzi, Ashish Sabharwal, Oren Etzioni, and Siena Dumas Ang. 2015. Parsing Algebraic Word Problems into Equations. Transactions of the Association for Computational Linguistics, 3:585–597.
[17] Talmor, A., Herzig, J., Lourie, N., & Berant, J. (2018). CommonsenseQA: A Question Answering Challenge Targeting Commonsense Knowledge. ArXiv. /abs/1811.00937
[18] Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2021. Transformer Feed-Forward Layers Are Key-Value Memories. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 5484–5495, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)