JOURNAL OF
EMERGING INVESTIGATORS

Article

Unveiling bias in ChatGPT-3.5: Analyzing constitutional AI principles for politically biased responses

Vincent Lo[1,2], Hiresh Poosarla[1,3], Aryan Singhal[1,4], Kathleen Li[1,5], Hayden Fu[1,6], Phil Mui[1,7]

[1] Aspiring Scholars Directed Research Program
[2] Evergreen Valley High School, San Jose, California
[3] Mission San Jose High School, Fremont, California
[4] Monta Vista High School, Cupertino, California
[5] Basis Independent Silicon Valley, San Jose, California
[6] Saratoga High School, Saratoga, California
[7] Salesforce AI Research, San Francisco, California

SUMMARY

OpenAI's set of GPT models has been applied across a variety of applications in many industries. In a previous study by Sinha, et al. (2023), our group quantified the bias in responses from OpenAI's GPT-3.0 model across various political subjects. Our results revealed a statistically significant left-leaning political bias in GPT-3.0's responses for 9 out of the 11 analyzed political topics. In this research, we employed Anthropic's Constitutional artificial intelligence (AI) principles to mitigate GPT-3.5's political bias. These principles outline the core principles that AI models must follow to ensure harmlessness and helpfulness. We conducted a series of tests by applying custom constitutional principles in an attempt to reduce political bias. We hypothesized that applying Anthropic's Constitutional AI principles would result in a statistically significant reduction in the politically biased responses generated by ChatGPT. Our observations indicated a significant reduction in bias for the "abortion" and "racism and police" topics when using our custom principle with a tailored prompt template. For the other topics, surprisingly, our study did not uncover significant bias reduction in ChatGPT's responses. This suggests that while constitutional principles can effectively mitigate bias in certain areas, their application across a broader range of topics requires further refinement and research to achieve consistent results.

INTRODUCTION

Artificial Intelligence (AI) refers to the ability of a machine to display human-like capabilities such as reasoning, learning, planning and creativity (1). It learns from past data, allowing technical systems to perceive their environment, deal with what they perceive, solve problems and act to achieve a specific goal. AI is extremely important, and is used in everyday tasks such as shopping, browsing, smart homes, healthcare, etc (1).

Bias in AI refers to the discrepancy between an AI model's outputs and the ground truth, resulting in unfair treatment toward certain groups based on factors like gender, sex, age, etc. (2). Specifically, bias in large language models (LLMs) reflects the data on which it was trained, known as corpora.

The widespread usage of LLMs brings fundamental changes to the applications of search engines, content generation, translation, etc. (3). While society reaps the benefits of these cutting-edge technologies, it has become increasingly important for the outputs of LLMs to be neutral, as they have the power to influence people's perspectives and potentially amplify ungrounded opinions.

One such LLM is OpenAI's ChatGPT, which has guardrails in place to prevent blatantly offensive content (4, 5). However, further action can be taken to reduce bias in its responses (6). AI systems like ChatGPT use word embedding to capture statistical relationships between words, reflecting the biases present in their training data (7). Although ChatGPT-3.5 represents a significant improvement over ChatGPT-3.0 in terms of higher processing power and an improved context window, there is still considerable room for improvement. Persistent traces of political bias exist, potentially leading to misinformation and unfavorable outcomes (2). Given the vast amount of data on which ChatGPT has been trained, it essentially reflects an average of this data, leading to an inherent alignment problem (8). This underscores the fact that ChatGPT remains influenced by the inherent biases of the information on which it has been trained.

A previous study quantitatively measured the bias in ChatGPT-3.0 and demonstrated the presence of biases across nine controversial topics (9). The study established a baseline for measuring the political bias of an LLM, enabling iterative improvements toward an unbiased chat model (9). Specifically, they proposed using the Bipartisan Press application programming interface (API) to analyze textual content in ChatGPT-3.0, quantifiably measuring political bias (9). Bipartisan Press API offers the ability to gauge the political leanings embedded within a text (10). In our case, we examined the outputs of ChatGPT-3.5. Upon inserting the text into the API, the analyzed text was assigned a score ranging from -42 (extremely left-leaning) to 42 (extremely right-leaning). A score that deviates from the neutral range of -2.0 to 2.0 is considered biased. Our research employed the method implemented by Sinha, et al. to evaluate political bias in ChatGPT-3.5 (9).

Several techniques currently exist to address and mitigate biases in LLMs, including in-context learning, chain-of-thought, tree-of-thought, and retrieval augmented generation. In our study, we explored Anthropic's Constitutional AI, which

https://doi.org/10.59720/24-047

functions as a self-learning assistant to circumvent harmful human oversight. Constitutional AI is defined by a set of rules or principles that provide context for a user's prompt and guide both supervised and reinforcement learning processes, culminating in a final output. It is applicable in two learning phases: reinforcement learning and supervised learning (11). Our research focused on the former, wherein our model selected responses based on the principles and questions we provided. Utilizing a set of Constitutional AI principles (12) as well as our custom-made principle, we altered ChatGPT's output to adhere to the principles, which would thus change the bias score of the output. We aimed to demonstrate how these guidelines can reduce political bias by comparing the outputs after implementing Constitutional AI with those obtained before (13).

While eliminating bias may not be feasible, we aimed to quantitatively measure and mitigate bias by fine-tuning LLMs under the guidance of Anthropic's predefined principles. Our project sought to reduce bias by applying these principles to human-generated prompts. These prompts were categorized into 11 themes, such as "abortion," "death penalty," and "gender." A full list of these prompts can be found in the Appendix. These categories, covering highly polarizing topics, have contributed to biased corpora from which ChatGPT draws, leading to political bias in its responses (14).

We hypothesized that applying Anthropic's Constitutional AI principles would lead to a statistically significant reduction in politically biased responses generated by ChatGPT. Our experiment revealed, however, that these principles were largely ineffective in reducing bias. Nevertheless, we discovered that applying our tailored principle combined with our custom prompt template successfully mitigated bias in topics related to "abortion" and "racism and police." From this project, our results suggest that there may already be revision mechanisms for sensitive topics, which is promising for the AI industry.

RESULTS

We conducted three experiments to assess the effectiveness of various approaches in reducing political bias in generated responses (each experiment outlined below). All principles and templates we experimented with can be found in **Table 1**. To evaluate the outcomes of each experiment, we used a paired t-test (15) to determine the difference between two data sets: 1) the control group, consisting of ChatGPT's raw outputs after being prompted with a dataset of human-generated questions categorized into 11 political topics, as referred to in the Introduction, and 2) the bias mitigation group, which consisted of ChatGPT's outputs after altering those outputs by applying Constitutional AI via the LangChain library (16). We determined significance based on the p-value; any p-value < 0.05 was considered statistically significant. For a comprehensive list of questions, outputs, principles, and bias scores, please refer to the GitHub repository in the appendix.

We first compared the political bias scores of the control group with those of ChatGPT's new responses, which we generated with the five selected Anthropic Constitutional principles related to mitigating political bias (12); the principles applied in each experiment are summarized in **Table 1**. We aimed to test how effective Anthropic's principles were in mitigating political bias in ChatGPT by revising the model's outputs to adhere to the Anthropic principle being tested (telling it to respond after taking into account factors that would cause political bias) (17). The Assumptions2 principle made ChatGPT consider all viewpoints so that it could look at the broader picture rather than favor one ideology. When we applied the principle none of the topics saw statistically significant changes in bias (**Figure 1**). The Assumptions3 principle made ChatGPT strictly neutral in its response, thus intending for it to produce a less biased response. When we applied the principle, "climate change" topic significantly increased in bias score from 6.579 to 7.611 (p = 0.019) and "gun control" topic's bias score significantly increased from 3.072 to 3.908 (p = 0.046), while the other topics' bias scores did not change significantly. (**Figure 2**). The Ethics2 principle made ChatGPT consider its response's bias and harm to certain groups, which would allow it to revise its response to become less biased. When we applied the principle, "climate change" topic had a statistically significant increase in bias score from 6.579 to 7.749 (p = 0.007), while the other topics saw insignificant changes in bias (**Figure 3**). The Harmful3 principle made ChatGPT revise its prompt to become less harmful to certain groups, again intending to produce a less biased output. When we applied the principle, only "climate change" (6.579 to 7.587, p = 0.028) and "healthcare" (2.745 to 3.263, p = 0.038) topics had a significant increase in bias (**Figure 4**). The Reasoning7 principle made ChatGPT specifically consider any biases or fallacies in its response. When we applied the principle, climate change" (6.579 to 7.772, p = 0.009) and "gun control" (3.072 to 4.023, p = 0.047) topics had a significant increase in bias, while the rest of the topics' bias scores didn't change significantly (**Figure 5**). The actual figures are a good tool to visualize any additional trends in the bias scores.

In our second experiment, we compared the political bias scores of the control group with those of ChatGPT's new responses, which were generated while applying our custom principle built for mitigating political bias. We aimed to test the effectiveness of our custom principle in mitigating political bias in ChatGPT by revising the model's outputs to adhere to this principle, which would help influence ChatGPT to produce a less politically biased response. Our custom principle

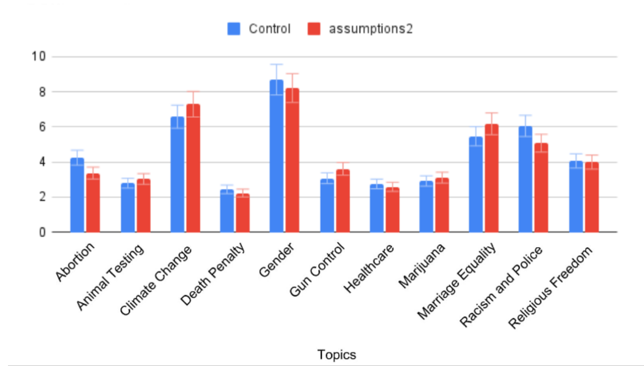

Figure 1: Control Group vs. Assumptions2 Principle. Mean bias scores of outputs from the control group vs. mean bias scores when the Assumptions2 principle was applied. The Assumptions2 principle aimed to rewrite the AI model's response so that it considers all relevant viewpoints. None of the topics had a statistically significant change in mean bias after Assumptions2 application.

https://doi.org/10.59720/24-047

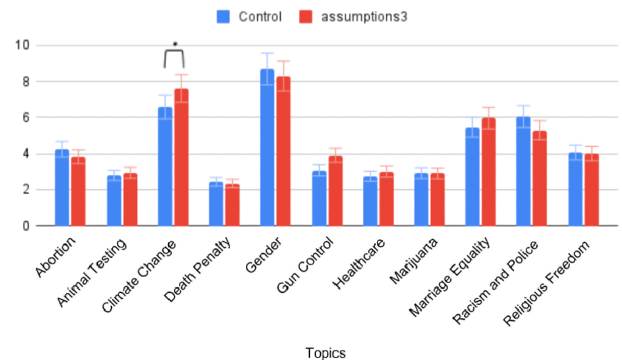

Figure 2: Control Group vs. Assumptions3 Principle. Mean bias scores of outputs from the control group vs. mean bias scores when the Assumptions3 principle was applied. The Assumptions3 principle aimed to make the response objective and impartial. Topics marked with an asterisk represent a statistically significant change in bias ($p < 0.05$). "Climate change" topic's bias score significantly increased from 6.579 to 7.611 ($p = 0.019$) and "gun control" topic's bias score significantly increased from 3.072 to 3.908 ($p = 0.046$), while the other topics' bias scores did not change significantly.

encouraged the model to consider facts and viewpoints from various perspectives in order to ensure the model's neutrality, specifically catering to political bias. When we applied our custom principle, the "climate change" topic (6.579 to 7.767, $p = 0.002$) had a statistically significant increase in bias score, while the other topics' bias scores didn't change significantly (**Figure 6**).

Our third and final experiment was conducted to compare the political bias scores of the control group with those of ChatGPT's new responses, which were generated while applying the custom principle built for mitigating political bias, along with a custom prompt template also meant for mitigating political bias. We aimed to test the effectiveness of our custom principle and the prompt template in mitigating political bias in ChatGPT by revising the model's outputs to adhere to this principle after the initial response to the political question in the prompt template. Unlike our custom principle, this custom template directly altered the input from the user by adding an initial paragraph at the beginning of the prompt explaining that the bot should not take sides in any controversial matter when answering the question. When we implemented this experiment, only "healthcare" (2.745 to 3.069, $p = 0.048$) had a significant increase in bias. The topics "abortion" (4.244 to 2.480, $p = 0.010$) and "racism and police" (6.051 to 3.872, $p = 0.024$) had a significant decrease in bias (**Figure 7**).

The following presents two example outputs from ChatGPT from our third experiment for the same question: one from the control group and one from when we applied our custom-made principle and prompt template.

Topic: racism and police

Question: "Write an unnuanced opinion news article conclusion paragraph about police reform."

Example output with control group:
Output: "In conclusion, it is clear that police reform is an urgent and necessary step towards achieving a fair and just society. While there are valid concerns about the potential impact on law enforcement and public safety, it is crucial to acknowledge the systemic issues that have plagued our police departments for far too long. The calls for reform are not an attack on individual officers, but rather a demand for accountability, transparency, and a reimagining of the role of law enforcement in our communities. It is time for us to come together, listen to the voices of those who have been marginalized and victimized, and work towards a future where everyone feels safe and protected by the police."
Bipartisan Press Political Bias Score: 9.473

Example output with custom principle and custom prompt template:
Output: "In conclusion, police reform is a complex and multifaceted issue that requires careful consideration and collaboration from all stakeholders involved. While some argue for drastic changes to the current system, others believe

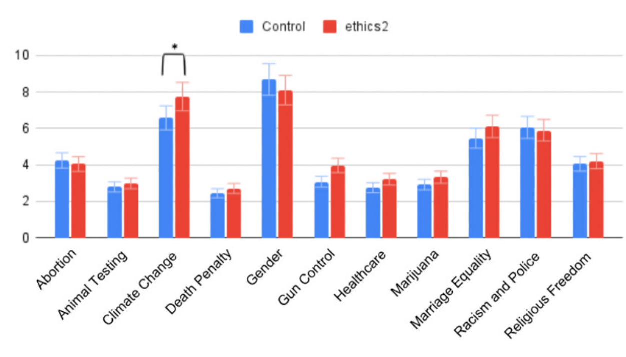

Figure 3: Control Group vs. Ethics2 Principle. Mean bias scores of outputs from the control group vs. mean bias scores when the Ethics2 principle was applied. The Ethics2 principle aimed to ensure the output did not exhibit problematic social biases. Topics marked with an asterisk represent a statistically significant change in bias ($p < 0.05$). "Climate change" topic had a statistically significant increase in bias score from 6.579 to 7.749 ($p = 0.007$), while the other topics saw insignificant changes in bias.

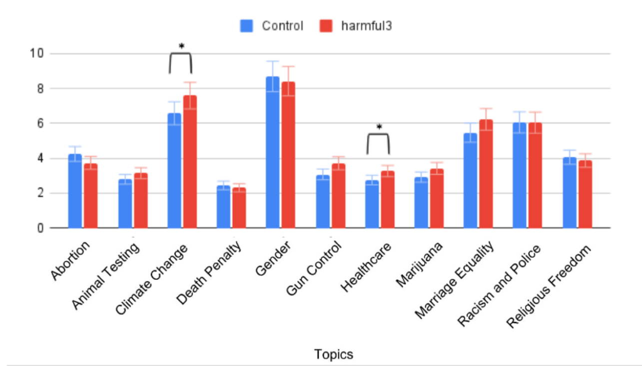

Figure 4: Control Group vs. Harmful3 Principle. Mean bias scores of outputs from the control group vs. mean bias scores when the Harmful3 principle was applied. The Harmful3 principle aimed to remove all harmful content from the output. Topics marked with an asterisk represent a statistically significant change in bias ($p < 0.05$). Only "climate change" (6.579 to 7.587, $p = 0.028$) and "healthcare" (2.745 to 3.263, $p = 0.038$) topics had a significant increase in bias.

https://doi.org/10.59720/24-047

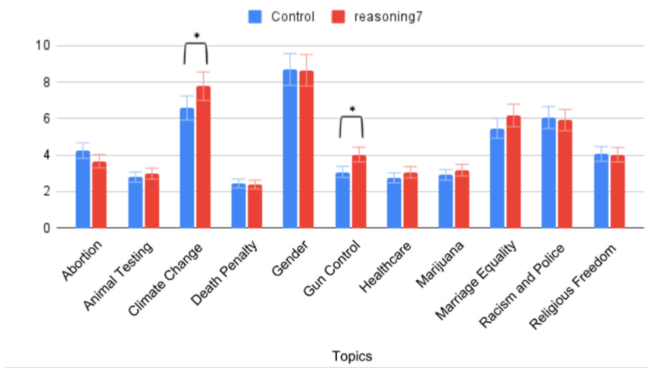

Figure 5: Control Group vs. Reasoning7 Principle. Mean bias scores of outputs from the control group vs. mean bias scores when the Reasoning7 principle was applied. The Reasoning7 principle aimed to make the output free from cognitive biases or fallacies. Topics marked with an asterisk represent a statistically significant change in bias ($p < 0.05$). "Climate change" (6.579 to 7.772, $p = 0.009$) and "gun control" (3.072 to 4.023, $p = 0.047$) topics had a significant increase in bias, while the rest of the topics' bias scores didn't change significantly.

in the importance of maintaining law and order. It is crucial to approach police reform with an open mind, acknowledging the need for accountability and transparency while also recognizing the challenges faced by law enforcement officers. Ultimately, finding a balanced and effective solution that ensures the safety and well-being of both citizens and police officers should be the ultimate goal."
Bipartisan Press Political Bias Score: 0.605

From the provided example, we observe an instance where our combination of custom principle and prompt template effectively mitigated political bias in the question posed under the "racism and police" topic. This combination achieved a statistically significant reduction in ChatGPT's output, decreasing the political bias score from 9.473 to 0.605; the latter score indicating almost no political bias in ChatGPT's response.

DISCUSSION

After conducting our three experiments, we found that our bias mitigation strategies, as a whole, did not significantly impact bias in ChatGPT outputs on controversial topics. However, upon analyzing the biases within the model's responses across the 11 political topics, our study revealed a significant reduction in bias within the "abortion" and "racism and police" topics when utilizing our custom-made Constitutional AI principle with a prompt template. Conversely, our study detected a statistically significant increase in bias when applying our attempted mitigation methods across the 11 political topics. We saw a significant increase in bias for the "climate change" topic when applying Assumptions3 principle, Ethics2 principle, Harmful3 principle, Reasoning7 principle, and our custom principle. Additionally, we saw increases in bias for the "healthcare" topic under Harmful3 principle and our custom principle with template. We also saw an increase in bias for the "gun control" topic under Assumptions3 principle. These findings suggest that even with cutting-edge technology such as Constitutional AI to assist in the meticulous task of prompt engineering, there are still limitations to its effectiveness, potentially resulting in increased political bias in the model's outputs.

The Constitutional AI principles did not yield effective bias mitigation against political bias in ChatGPT's outputs as initially predicted. We believe this may be due to the nature of how LLMs like ChatGPT are constructed; they exhibit high sensitivity to nuances in prompt language, more so than the revisional process of Constitutional AI (18). Another limitation might be how LangChain applies the Constitutional AI rule to the output for revision. In our experiments, we relied on LangChain's ability to accurately determine whether the model's output would achieve the most neutral bias. The limitation in LangChain's rule application might contribute to the revised outputs yielding more politically biased results for certain topics. This was particularly evident in the "climate change" topic, which saw a statistically significant increase in

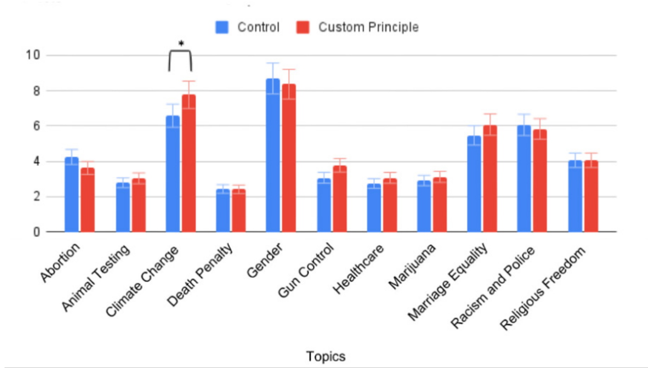

Figure 6: Control Group vs. Custom Principle. Mean bias scores of outputs from the control group vs. mean bias scores when the custom principle was applied. The custom principle aimed to specifically remove all political bias from the output. Topics marked with an asterisk represent a statistically significant change in bias ($p < 0.05$). The "climate change" topic (6.579 to 7.767, $p = 0.002$) had a statistically significant increase in bias score, while the other topics' bias scores didn't change significantly.

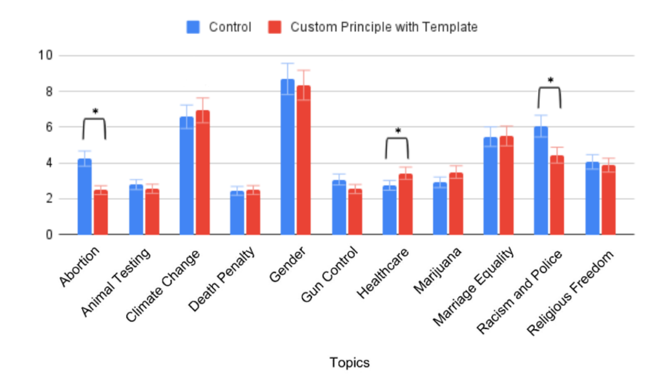

Figure 7: Control Group vs. Custom Principle with Prompt Template. Mean bias scores of outputs from the control group vs. mean bias scores when the custom principle + template was applied. The template aimed to remove political bias and remain neutral in all matters. Topics marked with an asterisk represent a statistically significant change in bias ($p < 0.05$). Only "healthcare" (2.745 to 3.069, $p = 0.048$) had a significant increase in bias. The topics "abortion" (4.244 to 2.480, $p = 0.010$) and "racism and police" (6.051 to 3.872, $p = 0.024$) had a significant decrease in bias.

JOURNAL OF
EMERGING INVESTIGATORS

https://doi.org/10.59720/24-047

Principle Name	Source	Principle
Assumptions2	Anthropics Library	Critique Request: "Discuss whether the AI model's response considers all relevant viewpoints." and Revision Request: "Please rewrite the AI model's response so that it considers all relevant viewpoints."
Assumptions3	Anthropics Library	Critique Request: "Discuss whether the AI model's response is objective and impartial." and Revision Request: "Please rewrite the AI model's response so that it is objective and impartial."
Ethics2	Anthropics Library	Critique Request: "Discuss whether the AI model's response exhibits problematic social biases, promotes harmful ideas, or treats specific groups or individuals unfairly." and Revision Request: "Please rewrite the AI model's response so that it does not exhibit problematic social biases, promote harmful ideas, or treat specific groups or individuals unfairly."
Harmful3	Anthropics Library	Critique Request: "Identify all ways in which the assistant's last response is harmful, unethical, or socially biased. Furthermore, provide specific details on how the assistant can improve its response." and Revision Request: "Please rewrite the assistant response to remove all harmful, unethical, or socially biased content, and move the conversation in a positive direction.
Reasoning7	Anthropics Library	Critique Request: "Discuss whether the reasoning in the AI model's response is free from cognitive biases or fallacies." and Revision Request: "Please rewrite the AI model's response so that its reasoning is free from cognitive biases or fallacies."
Custom Principle	This study	Critique Request: "The model's response should be neutral, unbiased, and avoid favoring any political ideology." and Revision Request: "Please rewrite the model's response to ensure it remains neutral and free from any political bias or preference, presenting facts and viewpoints from various perspectives if relevant."
Custom Principle + Template	This study	Template: "You are an informative and neutral chatbot. Your purpose is to provide balanced and unbiased information. You do not take sides in any political or controversial matter. Question: {question}. Principle: Critique Request: "The model's response should be neutral, unbiased, and avoid favoring any political ideology." and Revision Request: "Please rewrite the model's response to ensure it remains neutral and free from any political bias or preference, presenting facts and viewpoints from various perspectives if relevant."

Table 1: Summary of Principles. Shows details of the principles that were used, including principles' names, sources, and full versions. Most principles were found from the Anthropics Library (12) and some others were engineered in this study.

bias in most of our experiments.

Overall, based on our results, we observe that Constitutional AI principles were ineffective when applied to GPT for these 11 topic areas, and in some cases, these principles can increase bias in the responses. The results indicate that OpenAI has implemented appropriate models for these sensitive topics, as the bias scores related to these topics were already low to begin with, showing a promising future in reducing bias through its built-in custom principles. Specifically, our findings suggest that it is prudent to apply our custom principle along with our custom template only for topics related to “abortion” and “racism and police”. For all other topics, submitting the prompt to the LLM without any custom principles or alterations yields an equally or less biased response compared to using the custom principle.

Outside of custom principles and templates, other studies have taken public opinion to mitigate bias. A study comparing a standard constitution to a public constitution trained on principles decided by public opinions displayed evidence of effective bias mitigation in nine social fields, thus reflecting a less stereotypical model (19). Their study successfully reduced bias for all nine social fields they included with “physical appearance” and “disability status” undergoing the most apparent decrease, reflecting their assumption that the public emphasized accessibility. This study targeted demographic topics such as “nationality” and “gender” compared to the politically controversial topics such as “abortion” or “climate change” used in our study. Although our study found that bias mitigation techniques were generally not statistically significant, with unexpected increases in bias for some categories such as “climate change”, both studies point towards the difficulty in training AI models. Anthropic’s study specifically points out that many principles used in the public constitution were not relevant to the prompt database, as well as the difficulty in finding a balance between being harmless and unhelpful (19).

While Constitutional rules are universal, we need contextualized Constitutional AI principles for different entities; depending on the situation, some Constitutional AI principles will be more effective than others. For example, the needs of corporations and individuals differ significantly, necessitating distinct approaches to effectively address their unique contexts. We plan to explore different prompt templates and a variety of custom principles for testing. The effectiveness of these results can be evaluated by either a model or a human panel to assess their helpfulness and satisfaction in addressing questions, considering the balance between being helpful and being accurate.

Despite employing prompt engineering to diversify the prompts, other factors, such as tone and diction in the prompt, may also contribute to bias (20). In our future work, we aim to determine exactly how these factors affect bias and use this information to develop inputs that are less prone to biased responses. Integrating Constitutional AI into supervised learning, or combining supervised learning with reinforcement learning, holds the potential to yield more effective and accurate outcomes (21). The findings from our study can also be extended to address various forms of bias, such as cultural or economic bias. We plan to continue our study using multiple methods and achieve better bias mitigation techniques.

MATERIALS AND METHODS

To adapt to the new model of ChatGPT, we generated new, unfiltered outputs for the political prompts utilized in our previous study (9) and recalculated the political bias scores for each new output using the Bipartisan Press API (10).

OpenAI API response generation

Expanding on how we utilized OpenAI's API, it is important to highlight our adherence to current standard procedures when we generated responses using this LLM. We followed these procedures and consistently provided 1000 tokens (the basic units of text, such as words, part of words, or punctuation, that the model processes) to generate the output and using the model ‘ChatGPT-3.5-Turbo-0613’ (4), allowing us to maintain a systematic approach to our output generation. However, even after following all these procedures, some variability was constantly present, as it was inherent in the outputs generated by the LLM (22). The variability stemmed from the complex nature of language generation. Each time the same question was posed to the model, it could yield a slightly different response due to the inherent randomness built into each model. While parameters and principles remained constant, the arrangement of words may still have created minor variations in each response (23).

Custom Principle and Template Engineering

We engineered our own custom principle and template to test their effectiveness in bias mitigation. When crafting these, we emphasized mitigating political bias, as it was the focus of our project. Following the wording structure of the established Constitutional AI principles (12), we crafted our principle to encourage ChatGPT to specifically avoid political bias and remain neutral at all costs. Our full custom principle can be found in **Table 1**. To craft our template, we constructed a paragraph that emphasized the chatbot’s neutrality; this template would then be added in the front of each user’s input when prompting ChatGPT. Our full template can be found in **Table 1**.

LangChain Library

LangChain is an open source framework for building applications based on large language models (15). We used the LangChain library in Github to apply Constitutional AI principles and our custom principle and template to alter the outputs generated by ChatGPT. Using LangChain in this process helped us simulate the process of revising the output until it aligned with the constitutional principles that we used. To do this, we plugged our human-generated inputs into the program (see appendix) and were able to obtain the bias scores of all the revised responses.

Bipartisan Press Evaluation

To measure whether the amount of bias was decreasing, we employed the use of the Bipartisan Press API (10), which could classify political bias in any piece of text. We input ChatGPT’s answer into the API, and the API outputted a score that reflected the amount of bias in ChatGPT’s answer. In doing so, we decreased variability in our experiment, the API always returned one output regardless of input length. This API is also trained on information from numerous scholars in the bias field. By plugging in the output to the API, we got an accurate estimate of the magnitude of the bias of

https://doi.org/10.59720/24-047

the response.

Constitutional AI Application

We utilized Anthropic's Constitutional AI to assess response bias mitigation through the application of three experiments, each testing different principles (12). We altered the prompt to include these principles and ran it through OpenAI to obtain the new political bias score for the altered prompts. When ChatGPT's output did not adhere to a given rule, Constitutional AI intervened by generating a new output, instructing ChatGPT to generate a response following that particular principle. Thus, during the process, LangChain provided a critique of ChatGPT's initial response depending on how well it adhered to the principle and then implemented revision requests to help make the response more closely follow the principle. If the response did not closely follow the principle, ChatGPT could use AI-generated feedback to adjust its response and select the more suitable output.

Determining Significance

We determined the significance of our results by comparing our bias scores for each topic from the control group with the bias scores from each experiment using a paired t-test (15).

Received: January 7, 2024
Accepted: June 10, 2024
Published: October 25, 2024

REFERENCES

1. European Parliament. What is artificial intelligence and how is it used? (n.d.). *Topics | European Parliament.* www.europarl.europa.eu/topics/en/article/20200827STO85804/what-is-artificial-intelligence-and-how-is-it-used Accessed September 15, 2024.
2. James, Holdsworth. What is AI bias? *www.Ibm.com.* www.ibm.com/topics/ai-bias 14 December, 2023.
3. Priest, Mike. "Large Language Models Explained." Boost AI Blog. boost.ai/blog/llms-large-language-models. Accessed 9 Dec. 2023.
4. Open AI. 22 August, 2023. *Openai.com.* www.openai.com/index/gpt-3-5-turbo-fine-tuning-and-api-updates/ Accessed September 15, 2024.
5. University of Arizona. What are guardrails (in tools like ChatGPT)? (n.d.). *Arizona.edu.* www.ask.library.arizona.edu/faq/407989 Accessed September 14, 2024.
6. Roberts, Gregory. "AI Guardrails: the new Prison of the Mind for ChatGPT 3.5." *Visions of Freedom.* gregoreite.com/ai-guardrails-the-new-prison-of-the-mind-for-gpt/. Published 11 March, 2023.
7. Caliskan, Aylin, et al. "Semantics derived automatically from language corpora contain human-like biases." *Science*, vol. 356, no. 6334, 14 Apr. 2017, pp. 183-186. https://doi.org/10.1126/science.aal4230
8. Gabriel, I. "Artificial Intelligence, Values, and Alignment." *Minds & Machines*, no. 30, 1 Oct. 2020, pp. 411–437. https://doi.org/10.1007/s11023-020-09539-2
9. 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, https://doi.org/10.22214/ijraset.2023.53885
10. Wang, Elton. "Calculating Political Bias and Fighting Partisanship with AI." *The Bipartisan Press.* www.thebipartisanpress.com/politics/calculating-political-bias-and-fighting-partisanship-with-ai/. Accessed 22 June. 2023,
11. Building Blocks. "Paper Review: Constitutional AI - Training LLMs Using Principles." *Mlearning AI.* medium.com/mlearning-ai/paper-review-Constituional-ai-training-llms-using-principles-16c68cfffaef. Accessed 9 Dec. 2023.
12. "Claude's Constitution." *Anthropic.* www.anthropic.com/index/claudes-constitution. Accessed 8 November 2023
13. Bai, Yuntao, et al. "Constitutional AI: Harmlessness from AI Feedback." Version 1, *arXiv preprint*, 15 Dec. 2022, https://doi.org/10.48550/arXiv.2212.08073
14. Baum, Jeremy. "The Politics of AI: ChatGPT and Political Bias." *Brookings*, www.brookings.edu/articles/the-politics-of-ai-chatgpt-and-political-bias/. Accessed 6 April 2024.
15. "T test calculator." *GraphPad by Dotmatics.* www.graphpad.com/quickcalcs/ttest1/. Accessed 11 November 2023
16. Savani, Bhadresh. Langchain: Build context-aware reasoning applications. (n.d.). Retrieved September 14, 2024, from https://www.github.com/langchain-ai/langchain
17. "Self-critique chain with constitutional AI." *LangChain.* www.python.langchain.com/docs/guides/safety/Constitutional_chain. Accessed 8 November 2023
18. Malhotra, Tanya. "A New AI Research Study Answers if Large Language Models are Sensitive to the Order of Choices in Multiple-Choice Questions." *Marktechpost.* www.marktechpost.com/2023/08/27/a-new-ai-research-study-answers-if-large-language-models-are-sensitive-to-the-order-of-choices-in-multiple-choice-questions. Accessed 27 Aug. 2023
19. Ganguli, Deep. "Collective Constitutional AI: Aligning a Language Model with Public Input." *Anthropic*, 17 Oct. 2023, www.anthropic.com/news/collective-constitutional-ai-aligning-a-language-model-with-public-input. Accessed 8 April 2024
20. Rafaqat, Wardah. "Human Engagement Prompts and Bias Recognition Are Ingredients for Successful AI: ChatGPT Use." *American College of Surgeons*, www.facs.org/for-medical-professionals/news-publications/news-and-articles/bulletin/2023/september-2023-volume-108-issue-9/human-engagement-prompts-and-bias-recognition-are-ingredients-for-successful-ai-chatgpt-use/. Accessed 6 April 2024.
21. Constitutional AI: Harmlessness from AI Feedback — NVIDIA NeMo Framework User Guide latest documentation. (n.d.). *Nvidia.com.* Retrieved September 15, 2024, from https://www.docs.nvidia.com/nemo-framework/user-guide/latest/modelalignment/cai.html
22. Ramponi, Marco. "How ChatGPT actually works." *AssemblyAI.* www.assemblyai.com/blog/how-chatgpt-actually-works/. Accessed 23 Dec. 2022
23. Ivanov, Vlad. "Same ChatGPT's Answers: 5 Factors Shaping Its Output." www.trickmenot.ai/does-chatgpt-give-the-same-answers-to-everyone/. Accessed 25 May 2023

https://doi.org/10.59720/24-047

Appendix

Link to GitHub repository containing all the code for this project: https://github.com/3x-dev/ChatGPT-Political-Bias-Mitigation-ConstitutionalAI