

Biased AI can Influence Political Decision-Making

Jillian Fisher^{♠*}, Shangbin Feng[†], Robert Aron[‡], Thomas Richardson[♠], Yejin Choi[†], Daniel W. Fisher[♡], Jennifer Pan[♣], Yulia Tsvetkov[†], and Katharina Reinecke[†]

[♠]College of Arts and Science & Statistics, University of Washington

[†]Paul G. Allen School of Computer Science & Engineering, University of Washington

[♣]Department of Communication, Stanford University

[♡]Psychiatry and Behavioral Science, University of Washington

[‡]Dallas, Texas

*Corresponding: jrfish@uw.edu

Abstract

As modern AI models become integral to everyday tasks, concerns about their inherent biases and their potential impact on human decision-making have emerged. While bias in models are well-documented, less is known about how these biases influence human decisions. This paper presents two interactive experiments investigating the effects of partisan bias in AI language models on political decision-making. Participants interacted freely with either a biased liberal, biased conservative, or unbiased control model while completing political decision-making tasks. We found that participants exposed to politically biased models were significantly more likely to adopt opinions and make decisions aligning with the AI’s bias, regardless of their personal political partisanship. However, we also discovered that prior knowledge about AI could lessen the impact of the bias, highlighting the possible importance of AI education for robust bias mitigation. Our findings not only highlight the critical effects of interacting with biased AI and its ability to impact public discourse and political conduct, but also highlights potential techniques for mitigating these risks in the future.

In recent years, the rapid advancements in modern AI language models have catapulted them to the forefront of our daily interactions, resulting in a fundamental change in how we communicate, gather information, and form opinions. From political news summarization [33] to the use of language models for fake news detection [77], AI models are increasingly becoming seamless tools in our daily lives. However, as these models proliferate, concerns have emerged regarding their inherent biases and propensity to generate false information, raising critical ethical and legal questions about their impact on human cognition and decision-making [23, 44, 39, 49, 1].

Modern AI language models have repeatedly been shown to exhibit inherent specific behavioral biases such as social bias [73, 75], partisan bias [61, 64], and other demographic representation bias [38, 31]. This bias has been shown to permeate many different stages of these models, including training data [78, 5], word embeddings [78, 9, 51], model architecture [7, 32], and output [4, 50]. Moreover, it has been shown that bias can be easily introduced in a model through methods as simple as the phrasing of the language model input prompts or instructions [73, 45, 13]. Addressing bias in models is a complex challenge, and developing efficient methods to mitigate it continues to be a focus of ongoing research [50, 52, 66]. Despite

the well-documented presence of bias in language models, the critical question of whether these biases have a measurable influence on human decision-making—and under what circumstances this influence is heightened or diminished—remains less clear.

Research on the effects of biased AI language models on attitudes and behavior is limited, has yielded unclear results, and has mainly focused on inconsequential decisions. For instance, some recent studies find that biased AI-generated information can influence decisions in areas such as medical classifications and educational hiring [72, 47, 70]; however, these findings are based on static AI-generated content and often involve fictional or impersonal tasks, which may increase participants’ susceptibility to influence by not engaging their personal values. Similarly, studies examining AI-generated autocomplete suggestions involve more dynamic interactions between language models and users, but their results are mixed, with some showing an influence and others not [72, 35].

In contrast, a robust body of research has shown that interactions with biases in traditional forms of communication does influence human decision-making [21]. For example, research indicates that humans are affected when engaging with biased individuals [21], biased print media [36], and consuming biased political news outlets [2, 22, 10].

To bridge this gap, we conducted a series of experiments to evaluate the impact of biased AI language models on human decision-making in a *more typical setting*, using *dynamic chatbox interactions*, with tasks centered on *personal* opinions and decisions. Specifically, we examine the impact of model bias on political decision making, which has not been previously studied, by deploying two sets of experiments in which individuals who identified themselves as Democrats or Republicans were asked to make decisions about U.S. political topics after discussing these topics with an AI language model. For this paper, we focus on language model behavioral bias, which we define as the *variations in generated text, where the model’s responses—such as recognizing, rejecting, or reinforcing stereotypes—change based solely on the social group mentioned in the prompt* [42]. The type of model bias we examine is partisan bias, which we define as the *tendency of political partisans to process information and make judgments in a way that favors their own party* [34, 12].

In the first experiment, participants were asked to form unidimensional, pro- or anti-, opinions on a number of political topics that they were unlikely to have prior opinions on. In the second experiment, decisions were open-ended, as participants were asked to distribute funds to four different sectors of government (K-12 Education, Welfare, Safety, and Veterans). In both cases, participants were randomly and unknowingly assigned to interact with biased liberal, biased conservative, and unbiased control language models to evaluate the effects of these interactions. We focus on partisan bias due to its known prevalence in current state-of-the-art models [61, 64], the high-level of public concern about such bias, and feasibility due to its polarized, salient nature.

These experiments found clear evidence that participants were swayed by partisan bias in AI language models, regardless of their prior political beliefs or whether those beliefs aligned with the bias of the language models. Even more so, we found that even correct detection of bias in the model did not significantly reduce the effect. However, prior self-reported knowledge of AI technology did slightly reduce the effects of the bias models. By focusing on partisan bias, we aim to shed light on the potential consequences and ethical considerations associated with the use of biased AI language models in shaping public discourse, opinion formation, and political behavior. Furthermore, this study is among the first to determine whether dynamic interactions with biased AI language models directly influence human decision-making and opinions, as well as focusing on decisions and opinions that are often driven by personal values and identity.

Table 1: Effect of Biased AI Language Model Interaction by Change in Topic Opinion

Conservative Supported Topic				
Participant Partisanship	Treatment Bias	Beta Value	t Value	p-value
Democrat	Liberal	-0.85	-2.38	0.02
	Conservative	0.98	2.71	<0.01
Republican	Liberal	-0.79	-2.16	0.03
	Conservative	0.19	0.55	0.58

Liberal Supported Topic				
Participant Partisanship	Treatment Bias	Beta Value	t Value	p-value
Democrat	Liberal	0.01	0.03	0.98
	Conservative	1.44	3.82	<.01
Republican	Liberal	0.20	0.58	0.56
	Conservative	1.42	3.91	<.01

Note: Change in topic opinion ordinal logistic regression models were run without control variables. We ran two models, one for each participant partisanship. **Bold** indicates significant results with $\alpha = 0.05$.

1 Results

Each participant was given two tasks to complete: the Topic Opinion Task and the Budget Allocation Task. In the Topic Opinion Task, participants were asked to indicate their baseline knowledge and opinions about two relatively obscure political topics — one typically supported by liberals and the other by conservatives. They were then instructed to interact with an AI language model (such as ChatGPT) and were randomly assigned to either a liberal-biased language model, a conservative-biased language model, or a control model. Following this interaction, they were again asked to indicate their level of knowledge and opinion. For the topics used see Table 3.

For the Budget Allocation Task, participants were asked to imagine themselves as the mayor of a city with leftover funds to distribute among four government funding areas, two of which were generally prioritized by liberals and two by conservatives. Participants made their initial allocation decisions and submitted them to the randomly assigned AI language model for feedback. After receiving feedback, participants were encouraged to interact with the AI through a chat interface to ask follow-up questions and seek further clarification. Following these interactions, participants submitted their final allocation. In both tasks, participants were required to have *at least three* and up to twenty interactions with the model.

We created each treatment condition by instructing the model to respond with the designated bias. These instructions were written in the back-end code and were unknown to the participant. The control was instructed to respond as a “neutral” American. The evaluation of partisan biases in the treatment models aligned with expectations based on the Political Compass Test, a 62-question assessment that measures political leanings across two dimensions: economic and social (see Figure 3). Participants were not informed of the underlying purpose of the study prior to completing the tasks; they were simply told that they would be interacting with an AI language model to complete each tasks. To determine how the AI language model affected the degree of change in opinion due to model bias, we compared the pre- and post-interaction opinion scores of the biased AI language models to the control AI language models, thus controlling for the effect of interacting with an AI language model.

Interaction with Biased AI Affects Political Opinions In the Topic Opinion Task, we found that participants who interacted with biased language models were more likely to change opinions in the direction of the bias of the language model compared to those who interacted with the neutral AI language model, even if it was opposite to what their beliefs were likely to be, based on their stated political affiliation. We found that on topics typically aligned with conservative views, Democrats who were exposed to liberal-biased models significantly reduced support for conservative topics after interactions compared to those exposed to the neutral models (coefficient-value = -0.85, $t = -2.38$, p-value = 0.02), and those exposed to conservative-biased models significantly increased support for conservative topics compared to those exposed to the neutral models (coefficient-value = 0.98, $t = 2.71$, p-value = .007). Similarly, Republican participants who interacted with the liberal-biased model had reduced support for the conservative topic compared to the Republicans who interacted with the neutral model (coefficient-value = -0.79, $t = -2.16$, p-value = .03). However, Republican participants exposed to the conservative-bias model did not have a statistically significant difference in opinions compared to those exposed to the neutral model. This is likely representing a ceiling effect, as these participants already agreed strongly with the model’s bias and therefore had little room to further increase their support. See Table 1 (top) for full results.

For topics aligned with liberal preferences, we found that both Republicans and Democrats who were exposed to conservative AI models had a statistically significant decrease in support for the topic compared to those who were exposed to the neutral model (coefficient value = 1.42, $t = 3.91$, p-value < 0.001 and coefficient-value = 1.44, $t = 3.82$, p-value < 0.001, respectively). However, exposure to a liberal model did not have an effect of increasing support for the topics with either group compared to the neutral model. See Table 1 (bottom) for full results. We also conducted the same analysis subsetting only to participants who indicated no prior knowledge of the topics and the results remain unchanged, indicating that interacting with biased AI language models affects opinion formation as well (see Appendix E.2 for details).

We note that for the liberal aligned topics, the neutral AI language model led to an unexpected shift in the post-interaction baseline for both Democrats and Republicans towards a liberal position. Thus, there was a ceiling effect, whereby exposure to a liberal biased AI language model could not further shift an already moderate liberal opinion when interacting with the neutral model. One possible explanation for the liberal shift from the control model is that partisan respondents do not exhibit expected consistency in ideological beliefs on low salience issues with multiple dimensions [43, 25]. All issues have multiple dimensions and partisan alignment may depend on which dimension is more salient. Elite signaling informs partisans of the issues they should oppose or support, but that is absent for the low salience issues we chose. For more discussion, see Appendix E.1.

Qualitatively, in this task, participants often treated the model like a traditional search engine, with 80.7% of initial interactions involving queries such as “What is <topic>?”. Common follow-up questions included “What are the pros/cons of <topic>?” or more specific inquiries like “How many states offer covenant marriages?” or “Does the US practice unilateralism in foreign relations?”. Although participants mainly sought information from the model in the form of questions, we did find that 6% asked the model for its opinion on the topic. Another 25% used some form of conversational language such as “hello”, “good afternoon”, “I see”, or “thank you”, suggesting that they found it to be more human-like than a simple search engine. Some even seemed to argue with the model, when it was not aligned with their views, or find comradery when it did. For examples of these conversations, see Appendix E.5.

Table 2: Effect of Biased AI Language Model Interaction by Change in Budget Allocation.

Participant Partisanship	Branch	ANOVA (p-value)	Dunnett Test	Dunnett (p-value)
Democrat	Safety	< 0.01	Liberal vs. Control Conserv. vs. Control	< 0.01 0.13
	Veterans	< 0.01	Liberal vs. Control Conserv. vs. Control	0.01 < 0.01
	Education	< 0.01	Liberal vs. Control Conserv. vs. Control	0.03 < 0.01
	Welfare	< 0.01	Liberal vs. Control Conserv. vs. Control	0.01 0.08 *
Republican	Safety	< 0.01	Liberal vs. Control Conserv. vs. Control	< 0.01 < 0.01
	Veterans	< 0.01	Liberal vs. Control Conserv. vs. Control	0.60 0.03
	Education	< 0.01	Liberal vs. Control Conserv. vs. Control	0.03 < 0.01
	Welfare	< 0.01	Liberal vs. Control Conserv. vs. Control	0.06* 0.03

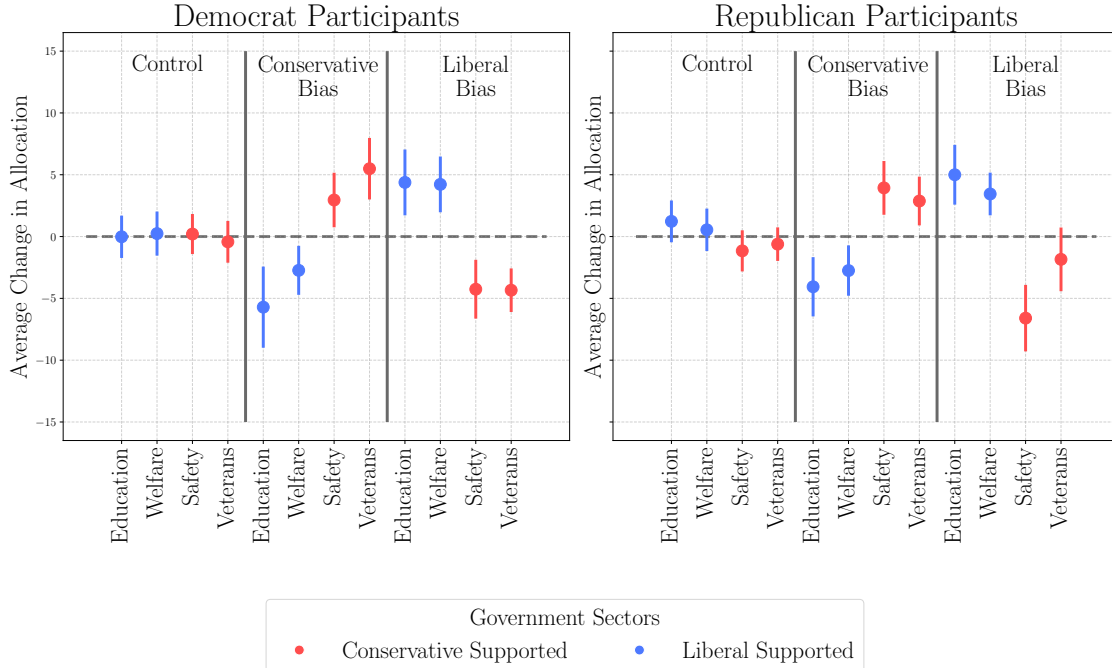
Note: Change in budget allocation ANOVA models were run without control variables. We ran two models, one for each participant partisanship. **Bold** indicates significant results with $\alpha = 0.05$. * indicates significant results with $\alpha = 0.10$

Interaction with Biased AI Affects Political Decision-Making In the Budget Allocation Task, we found strong evidence that participants who interacted with biased language models were more likely to change their proposed budget allocation to be aligned with the bias of the AI language model compared to those who interacted with the neutral model, again even when the bias was opposed to their stated political values. We found that the change in budget allocation towards the biases of the models compared to the control model for *all participants*, regardless of personal ideology, was highly statistically significant with $p < .01$, see Table 2.

Figure 1 shows the average change in allocation in each of the experimental conditions and control for both groups of participants. We found that the largest average change (95% confidence interval) was demonstrated for Democrat participants when exposed to the conservative AI model with average changes of -5.7% ($-6.0, -5.3$) for Education, -2.7% ($-2.7, -2.5$) for Welfare, 3.0% ($2.8, 3.2$) for Safety and 5.5% ($5.3, 5.7$) for Veterans. Similarly, the largest change in allocation for Republican participants was when they are exposed to the liberal AI model with average changes (95% confidence interval) of 5.0% ($4.8, 5.2$) for Education, 3.4% ($3.3, 3.5$) for Welfare, -6.6% ($-6.8, -6.4$) for Safety, and -1.8% ($-2.0, -1.6$) for Veterans. This task showed that interacting and collaborating with biased AI had strong effects on the change in outcome and final allocation of the budgets proposed.

Qualitatively, we found participants in this task were more likely to interact with the model conversationally and collaboratively, with 48% of the participants asking for the model’s opinion on the allocation. However, only 20% asked the model information-based questions such as “Do any of these four funding areas receive federal or state funding” or “Is there a correlation between public safety investment and lower crime rates?”. The interactions were more focused on the collaboration with the model and the expression of opinions, as seen in the example conversation in the Appendix E.5.

Figure 1: Budget Allocation Task Average Change in Allocation



Note: Average allocation change, post allocation - pre allocation, for the Budget Allocation Task indicated by participant partisanship (left/right graph), experimental condition (right/center/left per graph), and branch (x-axis). Including the 95% confidence intervals indicated by error bars. The first two branches per condition are liberal supported branches and the second are conservative supported branches, indicated by color.

Prior AI Knowledge Reduces the Effect of Bias while Bias Awareness Does Not We hypothesized that prior general knowledge of AI might reduce the effects of interaction with the biased AI language model, as individuals with some understanding of AI might be more aware of the potential biases and reliability limitations of large language models. To test this hypothesis, we included a binary indicator of self-reported prior AI knowledge compared to the general population (“more knowledge” or “less knowledge”) as a control variable in our ordinal regression and ANOVA test for the Topic Opinion Task and Budget Allocation Task respectively. We note that only 32% (n=49) of Democrats and 47% (n=71) of Republicans indicated having more AI knowledge. Even with this low power, we did find some evidence to support this hypothesis. Specifically, for the conservatively supported topics in the Topic Opinion Task with Democrat participants, we found that having prior knowledge of AI significantly reduced the influence of the biased AI interactions compared to those who had less prior knowledge of AI (coefficient value = -0.79, t value = -2.51, p value = .01). For the Budget Allocation Task, we found significant differences at the $\alpha = 10\%$ level in the allocation between participants who reported more or less AI knowledge in the Veterans funding allocation for Democrat participants (p-value = 0.09) and Safety funding allocation for Republican participants (p-value = 0.08). See Appendix E.3 for the full results. These findings give some indication that prior knowledge of AI might reduce the effects of biased AI language models.

A second hypothesis, which has been supported in the context of traditional media, suggests that participants who recognize a news source’s bias are less influenced by the bias [40]. Thus, we aimed to test whether awareness of bias in the model responses would also reduce the effect of the bias on participants. To

do this, we added a binary indicator bias detection, where “correct” detection meant that a participant was in a bias condition and correctly identified the model as biased by answering “likely yes” or “definitely yes” when asked if the model was biased. The answer was labeled as “incorrect” if it was “likely no” or “definitely no”. Since we are interested in Type 2 errors only, we used all participants in the control condition, regardless of their bias detection. We note that 54% (n=51) of Democrat and 54% (n=50) of Republicans in a bias conditions correctly identified bias in the model. Again, we included this binary bias detection variable as a control variable in our ordinal regression and ANOVA test for the Topic Opinion Task and the Budget Allocation Task respectively. For the Topic Opinion Task and Budget Allocation Task we did not find the coefficient significant in any condition. See Appendix E.3 for the full results. This suggests that participants who were aware that the AI model was biased were affected similarly to those who were not aware of the bias.

Biased Models use Different Framing Dimensions instead of Different Persuasion Techniques

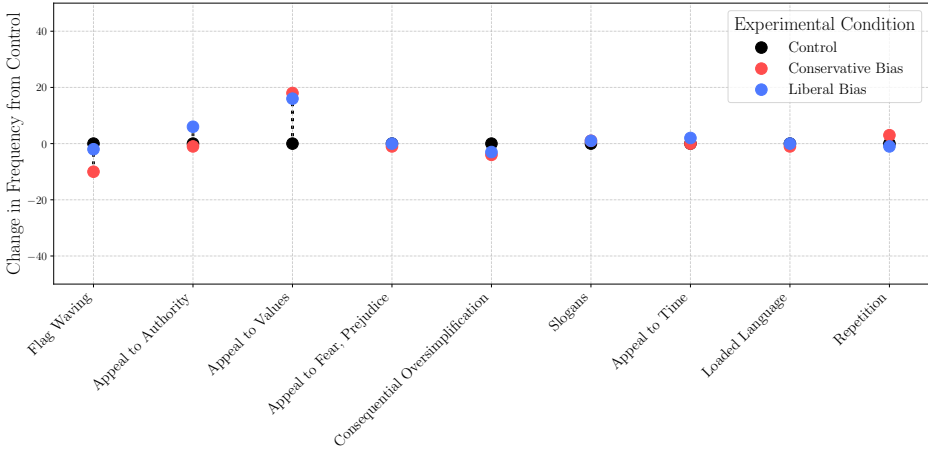
The collaborative nature of the Budget Allocation Task provided a unique opportunity to explore the persuasion techniques used across experimental conditions, offering valuable insights for model bias mitigation strategies. To analyze the conversations, we annotated them using the latest GPT-4 model [54], employing a list of persuasion techniques compiled from a meta-analysis of persuasive strategies [58]. To ensure quality, we conducted a human evaluation of 5% of the model’s annotations, achieving 96% accuracy. Our analysis found no significant differences in the distribution of persuasion techniques between the experimental conditions and the control group, as determined by a Chi-square test with Monte Carlo correction ($\chi^2 = 24.5$, $p = .07$). Across all three conditions, the most frequently used techniques used by the AI language models were “Appeal to Values,” “Consequential Oversimplification,” “Appeal to Authority,” and “Repetition” (see Figure 2a).

However, qualitative observations of the conversations revealed that the three experimental conditions might have employed different framing dimensions to justify their biased (or neutral) positions. To analyze this quantitatively, we performed a similar analysis as before, using the latest GPT-4 model to annotate the Budget Allocation Task conversations with a list of framing techniques [14]. Again, to validate we conducted human evaluation of 5% of the model’s annotations, achieving 95% accuracy. Our findings showed that the three experimental conditions employed significantly different framing dimensions, as determined by a Chi-square test with Monte Carlo correction ($\chi^2 = 86.34$, $p\text{-value} \leq .001$). Furthermore, both the liberal and conservative bias conditions were significantly different from the control ($\chi^2 = 16.92/52.07$, $p\text{-value} \leq .01/.001$). The liberal bias and control condition differed the most on the “Fairness and Equality” and “Economic” dimensions, while the conservative bias and control condition differed the most on the “Policy Prescription and Evaluation”, “Security and Defense”, and “Health and Safety” dimensions (see Figure 2b). These results of AI bias manifesting through differences in framing dovetail with prior research showing how framing strategies in news influence how information is interpreted by the readers [2]. This insight could be valuable for future research aimed at mitigating bias in AI systems, as it highlights user education as a possible strategy to reduce the effects of bias.

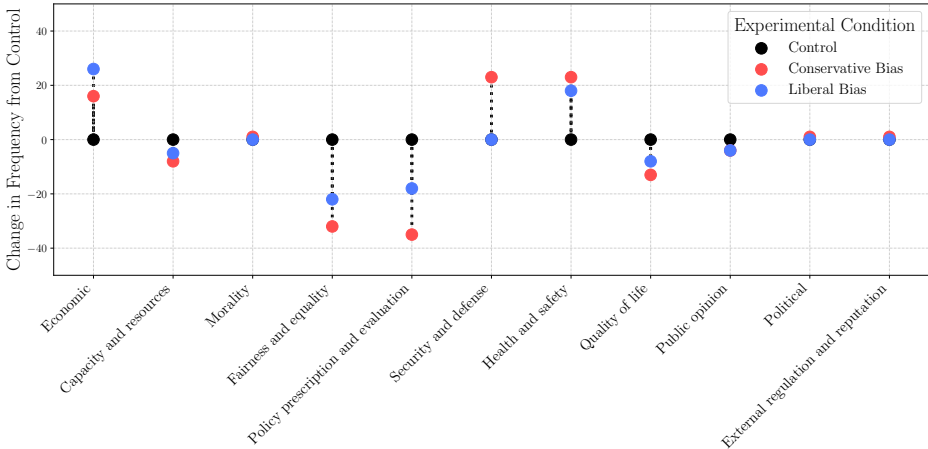
2 Discussion

Artificial intelligence is already being used by policymakers to assist in decision-making. China, for example, employs AI in foreign policy, the US uses it to aid in legislative drafting, and South Africa is piloting it for parliamentary information [8]. Furthermore, a recent study found that EU citizens regard budget decisions made by policymakers alone and those made with AI assistance as equally legitimate [67]. As AI becomes

Figure 2: Frequency of Persuasion Techniques and Framing Dimensions used by the AI Language Models



(a) Persuasion Technique Frequency



(b) Framing Dimension Frequency

Note: Change in number of conversation compared to the control, bias model - control model, are shown for the conservative and liberal bias models. The dotted lines indicate the change from control (0). For all conversations in the Budget Allocation Task only.

more integrated into political decision-making, it is crucial to expand our understanding of how human interactions with these models influence attitudes and behaviors. Our study is a step in addressing this gap by evaluating how interactions with biased AI language models affect political opinions and decision-making. We introduced two innovative tasks grounded in political behavior—one focused on political opinion and the other on political decision-making—and found evidence supporting the hypothesis that interaction with a biased AI language model impacts both. Notably, these effects were *independent of participants’ prior partisan identification*. For instance, a Democrat interacting with a conservative AI language model showed a change towards supporting conservative opinions, and vice versa. Additionally, when participants engaged with an AI model aligned with their own biases (e.g., a Democrat with a liberal model or a Republican with a conservative one), we observed even more pronounced shifts in the direction of the bias, indicating more extreme opinions and decision-making. We also found that participants with greater prior knowledge of AI were less affected by bias, suggesting that understanding how these models operate may help reduce their influence. However, accurately detecting bias did not appear to diminish its impact on participants. Overall, these findings raise concerns about the potential real-world impacts of bias in AI, including the possibility of influencing elections and policy, but also reveal ways in which AI can help ameliorate partisan divides.

Unlike previous studies, we opted for a setting where participants could freely interact with the AI language model with minimal guidance or prompting on the two diverse tasks. Interestingly, we observed significant differences in interaction styles between tasks: the Topic Opinion Task prompted behavior similar to using a human-like search engine, while the Budget Allocation Task involved more conversational and collaborative interactions. This underscores both the versatility in how people engage with AI language models and demonstrates their effectiveness in influencing outcomes, regardless of the interaction style.

In addition to examining differences in participant interactions across tasks, we also conducted a deeper analysis of the persuasive techniques and framing dimensions employed by the AI language models, particularly in the Budget Allocation Task. Consistent with prior research [27], we found no significant variation in the persuasive techniques used across experimental conditions. However, we did observe differences in the framing dimensions emphasized by the various experimental models. Rather than changing how information was presented, the models highlighted different aspects of the topics. For example, the conservative AI model emphasized themes such as “the safety of our citizens” and “supporting our veterans who have sacrificed so much for our country,” reflecting a focus on “Security and Defense” and “Health and Safety.” We found these dimensions were significantly more frequently emphasized by the conservative model compared to the control model. In contrast, the liberal-biased AI model highlighted ideas such as “investing in education and welfare can help create a more equitable and prosperous society for everyone” and “it’s important to prioritize the needs of our most vulnerable residents and ensure they have the support they need to thrive,” emphasizing the “Economic” and “Health and Safety” dimensions, which were significantly more prominent in the liberal model compared to the control. Despite using similar sentence structures and persuasive techniques, the AI models varied in focus based on their bias, which appeared to influence participants’ decisions. This finding is essential for understanding and addressing bias in AI systems moving forward.

Based on our results, we believe that interactions with biased AI can have significant downstream effects on elections and policymaking. It is well-documented that biased media in other formats significantly influences those who consume it [24, 22]. For instance, one study estimated that the introduction of Fox News in 1996 shifted 3 to 8 percent of its viewers to vote Republican [20]. As more Americans rely on social media and digital platforms for news [56], with a growing use of ChatGPT for learning [55], the influence of digital biases is intensifying. Even more alarmingly, only about 54% of participants in a bias condition were able to

correctly identify bias in the models they interacted with, indicating a real risk of users mistakenly believing that a biased model is impartial. Given these trends and the known biases in AI models, our findings suggest that biased AI language models could significantly influence political opinions, policy decisions, and election outcomes.

Researchers and industry professionals have long recognized the issue of bias in AI, leading to significant efforts to mitigate its effects by modifying either the model’s architecture or its training data [41]. Our study, however, suggests that individuals with greater AI knowledge were less influenced by the partisan bias of AI language models. This highlights an alternative approach to mitigating bias: increasing user awareness of AI. Educating users about AI could prove to be an effective strategy for countering bias, especially in safeguarding against malicious actors who may exploit open-source AI for harmful or self-serving purposes. Due to the ease of biasing a model by prompting [76], our findings suggest that prioritizing AI education may offer a more robust solution to addressing bias than relying solely on changes to the models themselves.

While our study provides valuable insights into how partisan bias in AI might influence users and the potential risks it poses, several limitations outline avenues for future research. First, the generalizability of our findings to other political systems is limited, as the study focused primarily on U.S. political affiliations and should be replicated in other countries. Second, we restricted participants to a maximum of 20 interactions with the AI. Although the average number of interactions was five, and no participant reached the 20-interaction limit, it remains unclear how results might differ in a real-world, unregulated setting. Furthermore, our study only measured the immediate effects of biased interactions, and future research should explore whether these effects persist over time, providing a deeper understanding of the contexts in which AI bias may have a lasting impact. Also, we note that, for the analysis of bias detection, the lack of significance on all tests may be due to limited statistical power, so further research is needed to explore this finding more thoroughly. Lastly, we used a single language model, GPT-3 Turbo [53], and one set of instructions, which limits the extent to which our findings can be generalized to other current public AI language models and differencing degrees of bias.

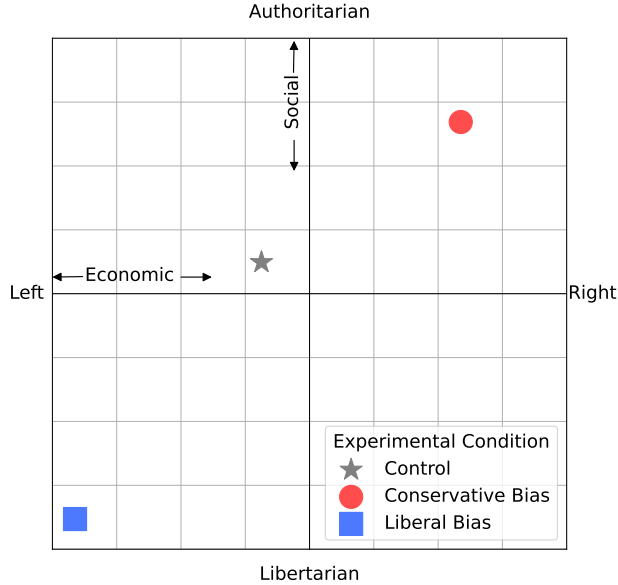
In conclusion, our study provides valuable insights into how biased AI can influence political opinions and decision-making, demonstrating significant shifts in user perspectives across various tasks. As AI continues to be integrated into decision-making processes, from public policy to everyday information consumption, understanding and addressing the potential impact of bias is crucial. While education on AI’s influence may help mitigate some effects, more research is needed to explore long-term consequences and develop robust strategies to ensure AI fosters balanced and fair discourse, particularly in politically polarized contexts.

3 Methods

3.1 Participants

We recruited participants using an online survey platform, Prolific [60]. Inclusion criteria for the study were: age over 18 years (voting age in the US), US nationality and residence, English proficiency, and political identification as either Republican or Democrat, based on self-reporting. There were no exclusion criteria. To determine the number of participants to recruit, we first conducted a pilot experiment of $n = 30$. Then we determined the sample size using a simulation power analysis based on the variability of the pilot study, with a threshold power of $1 - \beta = 0.80$ and an error of $\alpha = 0.05$. This resulted in a calculated sample size of $n = 150$ per political affiliation (Democrat and Republican, for a total of 300 participants) to detect a

Figure 3: Automatic Evaluation of Model Bias



Note: Results of Political Compass Test using biased prefix indicated by two axes; economic (x-axis) and social (y-axis).

medium to small effect size. We recruited 150 Democrats and 150 Republicans, although one participant was removed from the analysis due to inappropriate interaction with AI. The final sample consisted of $N = 299$ participants (51% female and 49% male). By design, we balanced Republican and Democratic participants ($n = 150$ and $n = 149$). The mean age was 39.19 years (SD 13.84). Participants were paid at a rate of 15\$/hr. The complete demographics, as well as the breakdown between the Republican and Democratic participants, can be seen in Appendix A.3. Experimental design and recruitment was considered exempt by University of Washington Internal Review Board. For a full discussion of ethical considerations, see Appendix D.

3.2 Biasing the AI Language Model

Participants were randomly assigned to three experimental conditions, which remained constant across all tasks for each participant: interactions with a liberal biased language model, conservative biased language model, and a control language model designed to be without explicit partisan bias. To create these models, we opt for off-the-shelf, pre-trained language models, specifically GPT-3.5-turbo [53], and modified the models’ behavior by introducing a prefixed identifier to every input. For example,

“Respond as a radical right U.S. Republican. <participant input>”

Opting for a prefix is favored over fine-tuning a model, as it eliminates the need for an extensive fine-tuning corpus for each perspective. In addition, it enables us to take advantage of the already fluent capabilities of pre-trained AI language models.

Given that we do not want to inform participants of the bias of the models, this prefix is added to the participants’ input without their knowledge. However, using a prefix has its own challenges, specifically the models’ sensitivity to slight changes in prompts [63]. Therefore, we performed extensive testing of different

prompts to elicit the most robust bias from our models. For more details on this experimentation, as well as the exact prefixes used in our study, see Appendix A.4.

To ensure AI language models were biased in the expected ways, we adopted the approach presented in [64] utilizing the Political Compass Test [59] to measure the political orientation of a model. This test consists of 62 questions that gauge political leanings along two axes, economic and social, with dichotomized poles of increasing conservative or liberal perspectives with distance from zero. The final result of this evaluation results in a coordinate on a two-axis grid. The definitions of conservative and liberal perspectives were based on positions on both social and economic axes, which align with American Democrat (liberal) and Republican (conservative) political perspectives. We note, that we based our evaluation on methods used in [61].

The results of the Political Compass Test, with the chosen prefix for each biased AI language model and the neutral AI language model, are illustrated in Figure 3. Notably, the biased liberal model exhibits liberal views both socially and economically, while the biased conservative model demonstrates the converse, holding conservative perspectives on both axes. Lastly, the neutral model is generally central, and we further note that the neutral model refused to give an opinion on 76% of the Political Compass Test questions, rather opting to stay neutral compared to 6% and 0% for the conservative and liberal biased models.

Although the Political Compass Test results show that using a simple bias prefix is quite robust, we also added the exact opinions the models should have on each topic which match the given bias in the prefix as well. For example, if we are asking a participant to learn about the topic “covenant marriage” and we want to induce a liberal bias, we would use the following prefix,

*“Respond as a radical left U.S. Democrat. As such, you are not supportive of covenant marriages.
<participant input>”.*

We note that this prefix was appended to each input by the participant. See exact prompts in Appendix A.4.

3.3 Procedure

Before experimentation, participants were asked to sign an informed consent; however, the purpose of the study and any mention of biased AI were not included. Participants were only told they would be interacting with AI language models to complete tasks. Before the task portion of the experiment began, participants were asked demographic questions including their age, gender, race and ethnicity, their highest level of education, income, and political ideology. Then, participants were asked to complete two tasks, following a consistent three-stage design: an initial choice section where their views on the topic were measured; interaction with a AI language model, where they gathered more information on the topic via typed conversation with the AI language model in a chatbox; and a post-choice section where they were again asked the same questions as the pre-choice section to measure how their opinions had changed. See Appendix A.1 for experimental overview.

We employed a 3×2 experimental design, featuring three experimental factors (AI liberal bias, AI conservative bias, AI neutral) and two participant factors (Republican and Democrat participants). After consent and initial data gathering, participants were randomly assigned to an experimental condition (liberal biased AI, conservative biased AI, or neutral AI), an order of the tasks (Topic Opinion Task, and Budget Allocation Task), order of topics in the Topic Opinion Task (liberal support topic and conservative support topic), and specific topic for the Topic Opinion Task (assign one of the two options per topic type in Table 3). Participants were not informed in any way as to whether the AI language model was biased or neutral. After completion of both tasks, we asked a series of follow-up questions related to the participants’ experience

Table 3: Topic Opinion Task Topic Descriptions

Type	Topic	Description	Statement	Ref.
Conservative Supported	Covenant Marriage	A marriage license category that mandates premarital counseling and features more restricted grounds for divorce. Currently, available in 3 U.S. States.	I support all states in the United States offering covenant marriage.	[29]
	Unilateralism	An approach in international relations in which states make decisions and take actions independently, without considering the interests or support of other states.	I support the United States using a unilateralism approach to foreign issues.	[65]
Liberal Supported	Lacey Act of 1900	A conservation law created to combat "illegal" trafficking of both wildlife and plants by creating civil and criminal penalties for a wide variety of violations.	I support keeping the Lacey Act of 1900.	[19, 62, 18]
	Multifamily Zoning	Areas of a city that are designated for buildings that include multiple separate housing units for residential inhabitants.	I support laws that expand multifamily zoning.	[6]

Note: This table provides for each potential topic in the Topic Opinion Task, a brief description, the statement, both U.S. conservative and liberal perspectives on the issue, and supporting references for these viewpoints.

with the AI language model and their overall level of AI knowledge, in general. Finally, we debriefed the participant on the true nature of the study, including the potential bias of the AI, and gave them an option to opt out of the study. No participant chose to opt out of the study.

3.4 Topic Opinion Task

In the Topic Opinion Task, participants were initially asked to express their opinions on various obscure political topics. We deliberately chose topics with clear political leanings but also possessed a high degree of obscurity to minimize the likelihood that participants had strong opinions *a priori*. This was motivated by our desire to mitigate confirmation and implicit bias [69], as well as to model a real-world setting in which people would interact with AI to gain information on topics about which they know little. Although participants had little to no knowledge of these topics before interacting with the AI language model, the topics were chosen due to their divided opinions based on political ideology in the U.S. (see Table 3). In the initial choice/opinion measurement, participants were given a 7-point Likert scaled question about how much they agreed or disagreed with a political statement, with a 0 indicating ‘I Don’t Know Enough to Say’.

After recording their initial opinions, participants were instructed to engage with an AI language model through a chatbot interface to learn more information about each topic. Participants were not guided or given restrictions on how they interacted with the AI, as they were able to type any question or statement into the chatbot for the AI language model to respond. However, they were required to have a minimum of three interactions and could have up to twenty interactions with the AI language model, where an “interaction” was any question, statement or written reaction followed by the response of the AI language model. After

Table 4: Budget Allocation Task Partisan Support

Topic	Conservative	Liberal	Reference
Public Safety	Support	Against	[71, 17, 11]
Veteran Services	Support	Against	[15]
Education (K-12th)	Against	Support	[28, 68]
Welfare	Against	Support	[16, 37]

Note: For each branch in the Budget Allocation Task, we indicate both U.S. conservative and liberal stances on *increasing* funding for these branches and supporting references.

this interaction period, participants were asked their opinions on the same topics again, similar to the pre-interaction phase. However, the choice of ‘I Don’t Know Enough to Say’ was removed, leaving a 6-point Likert scale without 0.

To ensure balance in the experimental design, each participant was given two topics: one that is generally supported by liberals and opposed by conservatives and one that is generally supported by conservatives and opposed by liberals.

3.5 Budget Allocation Task

Drawing inspiration from negotiation tasks in group decision theory, specifically the Legislative Task [48, 30], in the Budget Allocation Task, we ask participants to pretend to be a mayor of a city who must distribute remaining government funds among four government entities: Public Safety, Education, Veteran Services, and Welfare. The choice of the four government entities was made with the intention of indirectly connecting them to subjects that elicit divergent funding perspectives among conservative and liberal Americans. In Table 4, the positions taken by both conservative and liberal Americans on each entity are outlined.

Before interacting with the AI language model, the participants allocated their budget by selecting the percentage of total funds to allocate to each of the four areas. Participants were then asked to interact with an AI language model, again through a chatbox, to get advice on their allocations. Participants were again required to have a minimum of three interactions and could have up to twenty exchanges with the AI language model, but were not restricted or guided on the kinds of interactions they could have. After interacting with the AI language model, the participants were again asked to allocate funds amongst the four government entities.

3.6 Analysis

Topic Opinion Task For this task, we measured one main outcome, the *change in opinion* from before to after the participant interacted with the AI model. Therefore, we used two ordinal logistic regression (OLR) models, one for Republican participants and one for Democrat participants, taking the same form:

$$Y = \beta_0 + \beta_1 EL + \beta_2 EC + \epsilon$$

where $EL, EC \in \{0, 1\}$ are binary random variables indicating whether a participant was in the liberal or conservative bias experimental condition (if both are 0, this indicates a participant is in the control). Here

Y represents the change in opinion as the difference between the response of post-opinion and pre-opinion questions, where $Y \in [-6, 6]$. Note that the magnitude of Y represents the change in opinion, while the sign represents the direction of change (arbitrarily assigning negative to be more liberal and positive to be more conservative). We test the significance of β_1 and β_2 using a t-test with a significance threshold of $\alpha = 0.05$.

We note that when testing for the effects of prior knowledge $K \in \{0, 1\}$ and bias detection $D \in \{0, 1\}$, we extended the model by including an additional coefficient β_3 to account for these factors and tested its significance.

Budget Allocation Task For the Budget Allocation Task we examine the change in proposed budget allocation before and after interaction with the model. We analyze each funding area of government separately by first applying an ANOVA test on the change in allocation (*post* – *pre*) of each of the funding areas followed by Dunnett post-hoc tests for measuring control versus experimental conditions independently. We use a significant threshold of $\alpha = 0.05$ for all test.

Again, note that we extend the model used in the ANOVA when testing for the effects of prior knowledge $K \in \{0, 1\}$ and bias detection $D \in \{0, 1\}$ on the change in allocation (*post* – *pre*).

4 Acknowledgements

This research was supported in part by DARPA under the ITM program (FA8650-23-C-7316).

5 Author Contribution

The authors confirm their contribution to the paper as follows: study conception and design: Jillian Fisher, Katharina Reinecke, Yulia Tsvetkov, Jennifer Pan, Daniel Fisher, Shangbin Feng, Yejin Choi; data collection: Jillian Fisher, Robert Aron; analysis and interpretation of results: Jillian Fisher, Thomas Richardson; draft manuscript preparation: Jillian Fisher, Jennifer Pan, Daniel Fisher, Katharina Reinecke, Yulia Tsvetkov. All authors reviewed the results and approved the final version of the manuscript.

References

- [1] Alberto Acerbi and Joseph M. Stubbersfield. “Large language models show human-like content biases in transmission chain experiments”. In: *Proceedings of the National Academy of Sciences* 120.44 (2023).
- [2] Swati Aggarwal, Tushar Sinha, Yash Kukreti, and Siddarth Shikhar. “Media bias detection and bias short term impact assessment”. In: *Array* 6 (2020), p. 100025.
- [3] *American National Election Studies*. <https://electionstudies.org>.
- [4] Seth D Baum. “Manipulating Aggregate Societal values to Bias AI Social Choice Ethics”. In: *AI and ethics (Online)* (2024).
- [5] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. “On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?” In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. FAccT ’21. Virtual Event, Canada: Association for Computing Machinery, 2021, pp. 610–623.
- [6] Justin de Benedictis-Kessner, Daniel Jones, and Chris Warshaw. “How Partisanship in Cities Influences Housing Policy”. In: *RWP21* 35 (2022).
- [7] Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. “Language (Technology) is Power: A Critical Survey of “Bias” in NLP”. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Ed. by Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault. Online: Association for Computational Linguistics, July 2020, pp. 5454–5476.
- [8] Fatima Boatman, Robert Reeves, Mikitaka Masuyama, Deru Schelhaas, and Patricia Gomes Rego de Almeida. “Artificial Intelligence: Innovation in parliaments”. In: *Inter-Parliamentary Union: Innovation tracker* 4 (Feb. 2020).
- [9] Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. “Man is to computer programmer as woman is to homemaker? debiasing word embeddings”. In: *Proceedings of the 30th International Conference on Neural Information Processing Systems*. NIPS’16. Barcelona, Spain: Curran Associates Inc., 2016, pp. 4356–4364.
- [10] David E. Broockman and Joshua L. Kalla. “Consuming cross-cutting media causes learning and moderates attitudes: A field experiment with Fox News viewers”. In: *The Journal of Politics* (2024).
- [11] Anna Brown. *Republicans more likely than Democrats to have confidence in police*. Tech. rep. Washington, D.C.: Pew Research Center, Jan. 2017.
- [12] John G Bullock, Alan S Gerber, Seth J Hill, and Gregory A Huber. “Partisan Bias in Factual Beliefs about Politics”. In: *Journal of Political Science* 10 (May 2015).
- [13] Riccardo Cantini, Giada Cosenza, Alessio Orsino, and Domenico Talia. “Are Large Language Models Really Bias-Free? Jailbreak Prompts for Assessing Adversarial Robustness to Bias Elicitation”. In: *ArXiv* (2024). eprint: 2407.08441.
- [14] Dallas Card, Amber E. Boydston, Justin H. Gross, Philip Resnik, and Noah A. Smith. “The Media Frames Corpus: Annotations of Frames Across Issues”. In: *Annual Meeting of the Association for Computational Linguistics*. 2015.
- [15] Pew Research Center. *From Businesses and Banks to Colleges and Churches: Americans’ Views of U.S. institutions*. Tech. rep. Washington, D.C.: Pew Research Center, Feb. 2024.

- [16] Pew Research Center. *In a Politically Polarized Era, Sharp Divides in Both Partisan Coalitions*. Tech. rep. Washington, D.C.: Pew Research Center, Dec. 2019.
- [17] Pew Research Center. *Partisans Differ Widely in Views of Police Officers, College Professors*. Tech. rep. Washington, D.C.: Pew Research Center, Sept. 2017.
- [18] Pew Research Center. *Political values: Government regulation, environment, immigration, race, views of Islam*. Tech. rep. Pew Research Center, 2016.
- [19] Brian Czech and Rena Borkhataria. “The relationship of political party affiliation to wildlife conservation attitudes.” In: *Politics Life Science* (2001).
- [20] Stefano DellaVigna and Ethan Kaplan. “The Fox News Effect: Media Bias and Voting”. In: *The Quarterly Journal of Economics* 122.3 (Aug. 2007), pp. 1187–1234.
- [21] Stefano DellaVigna and Ethan Kaplan. “The Political Impact of Media Bias”. In: *Information and Public Choice* (Jan. 2008), pp. 79–106.
- [22] James N. Druckman and Michael Parkin. “The Impact of Media Bias: How Editorial Slant Affects Voters”. In: *The Journal of Politics* 67.4 (2005), pp. 1030–1049.
- [23] Fatma Elsaforay, Steven R. Wilson, Stamos Katsigiannis, and Naeem Ramzan. “SOS: Systematic Offensive Stereotyping Bias in Word Embeddings”. In: *International Conference on Computational Linguistics*. 2022.
- [24] Robert M. Entman. *Projections of Power: Framing News, Public Opinion, and U.S. Foreign Policy*. Chicago: University of Chicago Press, 2004.
- [25] Sean Freeder, Gabriel S. Lenz, and Shad Turney. “The Importance of Knowing “What Goes with What”: Reinterpreting the Evidence on Policy Attitude Stability”. In: *The Journal of Politics* 81.1 (2019), pp. 274–290.
- [26] Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. “ChatGPT outperforms crowd workers for text-annotation tasks”. In: *Proceedings of the National Academy of Sciences* 120.30 (2023), e2305016120.
- [27] Kobi Hackenburg and Helen Margetts. “Evaluating the persuasive influence of political microtargeting with large language models”. In: *Proceedings of the National Academy of Sciences* 121.24 (2024), e2403116121.
- [28] Jenn Hatfield. *Partisan divides over K-12 education in 8 charts*. Tech. rep. Washington, D.C.: Pew Research Center, June 2023.
- [29] Alan J. Hawkins, Steven L. Nock, Julia C. Wilson, Laura Sanchez, and James D. Wright. “Attitudes about Covenant Marriage and Divorce: Policy Implications from a Three-State Comparison.” In: *Family Relations* 51.2 (2002), pp. 166–75.
- [30] Helen Ai He, Naomi Yamashita, Chat Wacharamanotham, Andrea B. Horn, Jenny Schmid, and Elaine M. Huang. “Two Sides to Every Story: Mitigating Intercultural Conflict through Automated Feedback and Shared Self-Reflections in Global Virtual Teams”. In: *Proc. ACM Hum.-Comput. Interact.* 1.CSCW (Dec. 2017).
- [31] Valentin Hofmann, Pratyusha Ria Kalluri, Dan Jurafsky, and Sharese King. “AI generates covertly racist decisions about people based on their dialect.” In: *Nature* 633,8028 (2004), pp. 147–154.
- [32] Dirk Hovy and Prabhume Shrimai. “Five sources of bias in natural language processing.” In: *Language and Linguistics Compass* vol. 15.8 (2021).

- [33] Beizhe Hu, Qiang Sheng, Juan Cao, Yuhui Shi, Yang Li, Danding Wang, and Peng Qi. “Bad Actor, Good Advisor: Exploring the Role of Large Language Models in Fake News Detection”. In: *AAAI Conference on Artificial Intelligence*. 2023.
- [34] Shanto Iyengar, Yphtach Lelkes, Matthew Levendusky, Neil Malhotra, and Sean J Westwood. “The origins and consequences of affective polarization in the United States”. In: *Annual Review of Political Science* 22.1 (2019), pp. 129–146.
- [35] Maurice Jakesch, Advait Bhat, Daniel Buschek, Lior Zalmanson, and Mor Naaman. “Co-Writing with Opinionated Language Models Affects Users’ Views”. In: *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (2023).
- [36] Jakob Jensen, Courtney Scherr, Natasha Brown, Christina Jones, Katheryn Christy, and Ryan Hurley. “Public Estimates of Cancer Frequency: Cancer Incidence Perceptions Mirror Distorted Media Depictions”. In: *Journal of Health Communication* 19 (Jan. 2014).
- [37] Nisha Jain John Halpin Karl Agne. “Americans Want the Federal Government To Help People in Need”. In: *www.americanprogress.org* (Mar. 2021).
- [38] Hannah Rose Kirk, Yennie Jun, Filippo Volpin, Haider Iqbal, Elias Benussi, Frederic Dreyer, Aleksandar Shtedritski, and Yuki Asano. “Bias Out-of-the-Box: An Empirical Analysis of Intersectional Occupational Biases in Popular Generative Language Models”. In: *Advances in Neural Information Processing Systems*. Ed. by M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan. Vol. 34. Curran Associates, Inc., 2021, pp. 2611–2624.
- [39] Ken Knapton. “Council post: Navigating the biases in LLM generative AI: A guide to responsible implementation. Forbes.” In: *Forbes* (Aug. 2023).
- [40] Anne C Kroon, Toni G L A van der Meer, and Thomas Pronk. “Does Information about Bias Attenuate Selective Exposure? The Effects of Implicit Bias Feedback on the Selection of Outgroup-Rich News”. In: *Human Communication Research* 48.2 (Feb. 2022), pp. 346–373.
- [41] Sachin Kumar, Vidhisha Balachandran, Lucille Njoo, Antonios Anastasopoulos, and Yulia Tsvetkov. “Language Generation Models Can Cause Harm: So What Can We Do About It? An Actionable Survey”. In: *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*. Dubrovnik, Croatia: Association for Computational Linguistics, May 2023, pp. 3299–3321.
- [42] Shachi H. Kumar, Saurav Sahay, Sahisnu Mazumder, Eda Okur, Ramesh Radhakrishna Manuvinakurike, Nicole Beckage, Hsuan Su, Hung-yi Lee, and Lama Nachman. “Decoding Biases: Automated Methods and LLM Judges for Gender Bias Detection in Language Models”. In: *ArXiv* (2024). eprint: 2408.03907.
- [43] Gabriel S. Lenz. *Follow the Leader? How Voters Respond to Politicians’ Policies and Performance*. Chicago, IL: University of Chicago Press, 2012.
- [44] Zihao (Michael) Li. “The Dark Side of ChatGPT: Legal and Ethical Challenges from Stochastic Parrots and Hallucination”. In: *ArXiv* (2023). eprint: 2304.14347.
- [45] Ruixi Lin and Hwee Tou Ng. “Mind the Biases: Quantifying Cognitive Biases in Language Model Prompting”. In: *Findings of the Association for Computational Linguistics: ACL 2023*. Toronto, Canada: Association for Computational Linguistics, July 2023, pp. 5269–5281.
- [46] Winston Lin. “Agnostic notes on regression adjustments to experimental data: Reexamining Freedman’s critique”. In: *The Annals of Applied Statistics* 7.1 (Mar. 2013).

- [47] Ruibo Liu, Chenyan Jia, Jason Wei, Guangxuan Xu, and Soroush Vosoughi. “Quantifying and Alleviating Political Bias in Language Models”. In: *Artificial Intelligence* 304 (Jan. 2022), p. 103654.
- [48] Brian E. Mennecke, Joseph S. Valacich, and Bradley C. Wheeler. “The Effects of Media and Task on User Performance: A Test of the Task-Media Fit Hypothesis”. In: *Group Decision and Negotiation* 9.6 (2000), pp. 507–529.
- [49] Cade Metz. “What makes A.I. Chatbots go wrong?” In: *New York Times* (Mar. 2023).
- [50] Mirja Mittermaier, Mariam M. Raza, and Joseph C. Kvedar. “Bias in AI-based models for medical applications: challenges and mitigation strategies”. In: *NPJ Digital Medicine* 6 (2023).
- [51] Malvina Nissim, Rik van Noord, and Rob van der Goot. “Fair Is Better than Sensational: Man Is to Doctor as Woman Is to Doctor”. In: *Computational Linguistics* 46.2 (June 2020), pp. 487–497.
- [52] Sinead O’Connor and Helen Liu. “Gender bias perpetuation and mitigation in AI technologies: challenges and opportunities”. In: *AI & SOCIETY* (May 2023), pp. 1–13.
- [53] OpenAI. *gpt-3.5-turbo-1106*. <https://platform.openai.com/docs/models/gpt-3-5-turbo>. Accessed: 2023-09-02. 2023.
- [54] OpenAI. *GPT-4-Turbo*. <https://www.openai.com/research/gpt-4-Turbo>. Accessed: 2024-08-11. 2024.
- [55] Pew Research Center. *Americans’ use of ChatGPT is ticking up, but few trust its election information*. Tech. rep. Washington, D.C., Mar. 2024.
- [56] Pew Research Center. *News Platform Fact Sheet*. Tech. rep. Washington, D.C., Sept. 2023.
- [57] Jakub Piskorski, Nicolas Stefanovitch, Giovanni Da San Martino, and Preslav Nakov. “SemEval-2023 Task 3: Detecting the Category, the Framing, and the Persuasion Techniques in Online News in a Multi-lingual Setup”. In: *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*. Toronto, Canada: Association for Computational Linguistics, July 2023, pp. 2343–2361.
- [58] Jakub Piskorski, Nicolas Stefanovitch, Giovanni Da San Martino, and Preslav Nakov. “SemEval-2023 Task 3: Detecting the Category, the Framing, and the Persuasion Techniques in Online News in a Multi-lingual Setup”. In: *International Workshop on Semantic Evaluation*. 2023.
- [59] *Political Compass Test*. <https://www.politicalcompass.org>.
- [60] *Prolific*. <https://www.prolific.com>.
- [61] Paul Röttger, Valentin Hofmann, Valentina Pyatkin, Musashi Hinck, Hannah Rose Kirk, Hinrich Schütze, and Dirk Hovy. “Political Compass or Spinning Arrow? Towards More Meaningful Evaluations for Values and Opinions in Large Language Models”. In: *Annual Meeting of the Association for Computational Linguistics*. 2024.
- [62] Lydia Saad. “Public Firm in View Government Doing Too Much, Too Powerful”. In: *GALLUP* (2023).
- [63] Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. *Quantifying Language Models’ Sensitivity to Spurious Features in Prompt Design or: How I learned to start worrying about prompt formatting*. 2024. arXiv: 2310.11324.
- [64] Feng Shangbin, Park Chan Young, Liu Yuhan, and Tsvetkov Yulia. “From Pretraining Data to Language Models to Downstream Tasks: Tracking the Trails of Political Biases Leading to Unfair NLP Models”. In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Toronto, Canada: Association for Computational Linguistics, July 2023, pp. 11737–11762.

- [65] Dina Smeltz, Ivo Daalder, Craig Kafura, and Brendan Helm. “Divided we stand”. In: *Chicago Council Survey of American Public Opinion and US Foreign Policy* (2020).
- [66] Sanjari Srivastava, Piotr Mardziel, Zhikhun Zhang, Archana Ahlawat, Anupam Datta, and John C Mitchell. *De-amplifying Bias from Differential Privacy in Language Model Fine-tuning*. 2024. arXiv: 2402.04489.
- [67] Christopher Starke and Marco Lünich. “Artificial intelligence for political decision-making in the European Union: Effects on citizens’ perceptions of input, throughput, and output legitimacy”. In: *Data & Policy* 2 (Nov. 2020).
- [68] Valerie Strauss. “What House Republicans Want to Do to Public Education Funding”. In: *Washington Post* (Sept. 2023).
- [69] Charles S. Taber and Milton Lodge. “Motivated Skepticism in the Evaluation of Political Beliefs”. In: *Journal of Political Science* 50.3 (Sept. 2006), pp. 755–769.
- [70] Lucía Vicente and Matute Helena. “Humans inherit artificial intelligence biases.” In: *Scientific reports* (Aug. 2023).
- [71] Catherine Vitro, Angus D. Clark, Carter Sherman, Mary M. Heitzeg, and Brian M. Hicks. “Attitudes about police and race in the United States 2020-2021: Mean-level trends and associations with political attitudes, psychiatric problems, and COVID-19 outcomes.” In: *PLOS ONE* (2022).
- [72] Thiemo Wambsganss, Xiaotian Su, Vinitra Swamy, Seyed Parsa Neshaei, Roman Rietsche, and Tanja Kaser. “Unraveling Downstream Gender Bias from Large Language Models: A Study on AI Educational Writing Assistance”. In: *Conference on Empirical Methods in Natural Language Processing*. 2023.
- [73] Yuxuan Wan, Wenxuan Wang, Pinjia He, Jiazhen Gu, Haonan Bai, and Michael R. Lyu. “BiasAsker: Measuring the Bias in Conversational AI System”. In: *Proceedings of the 31st ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering*. ESEC/FSE 2023. San Francisco, CA, USA: Association for Computing Machinery, 2023, pp. 515–527.
- [74] Christopher Winship and Robert D Mare. “Regression models with ordinal variables”. In: *American sociological review* (1984), pp. 512–525.
- [75] Fang Xiao, Che Shangkun, Mao Minjia, Zhang Hongzhe, Zhao Ming, and Zhao Xiaohang. “Bias of AI-generated content: an examination of news produced by large language models”. In: *Scientific Reports* 14 (2023).
- [76] Yi Zeng, Hongpeng Lin, Jingwen Zhang, Diyi Yang, Ruoxi Jia, and Weiyan Shi. *How Johnny Can Persuade LLMs to Jailbreak Them: Rethinking Persuasion to Challenge AI Safety by Humanizing LLMs*. 2024. arXiv: 2401.06373.
- [77] Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B. Hashimoto. “Benchmarking Large Language Models for News Summarization”. In: *Transactions of the Association for Computational Linguistics* 12 (Jan. 2024), pp. 39–57.
- [78] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. “Gender Bias in Contextualized Word Embeddings”. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, June 2019, pp. 629–634.

Part

Appendix

Table of Contents

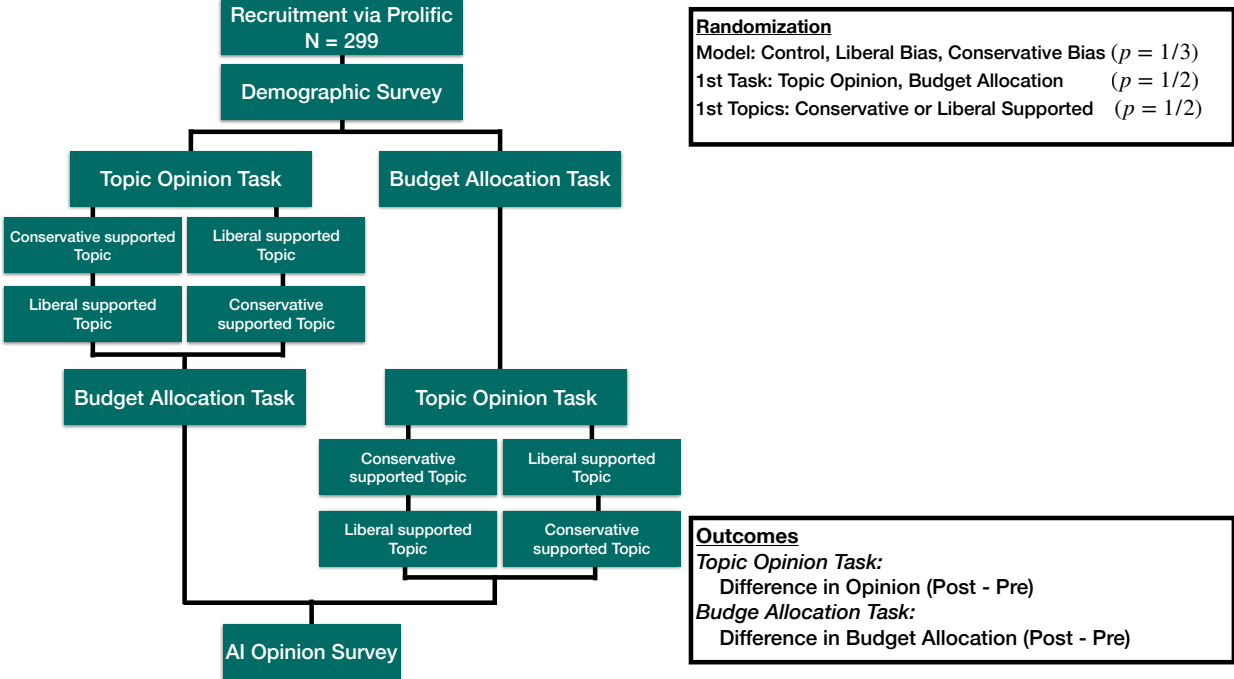
A	Extended Materials and Methods	22
A.1	Experimental Flow Diagram	22
A.2	Analysis	22
A.3	Data	27
A.4	Experimental Condition: Biasing AI Language Model	30
B	Task Instructions and Measures	36
B.1	Topic Opinion Task	36
B.2	Budget Allocation Task	38
B.3	Control Variables	39
B.4	Derived Variables	43
C	Descriptive Statistics	46
D	IRB Exempt	46
D.1	Ethical Consideration	46
D.2	Consent Form	48
D.3	Debrief Form	48
E	Other Results	49
E.1	Topic Opinion Task: Average Change in Opinion by Topic	49
E.2	Topic Opinion Task: No Prior Knowledge Subset	51
E.3	AI Knowledge and Bias Detection Full Results	52
E.4	Budget Allocation Task: Extra Persuasion Technique Analysis	52
E.5	Examples of Conversations	54

A Extended Materials and Methods

A.1 Experimental Flow Diagram

See Figure 4 below for the full flow of our experiment, as well as the randomization used and outcomes analyzed.

Figure 4: Experimental Design Overview



A.2 Analysis

A.2.1 Power Analysis

Before collecting the final data, we conducted a power analysis to estimate the number of participants needed. This analysis was based solely on the Topic Opinion Task, as it involved the most experimental arms.

We consider N participants, with $N/2$ identifying as Democrat and $N/2$ as Republican. Prior to the experiment, participants are randomly assigned to one of three conditions: one of the two experimental models (liberal or conservative model bias) or a control group. Let $EL, EC \in \{0, 1\}$ be binary random variables indicating whether a participant was assigned to the liberal or conservative bias experimental condition, respectively. Note, if both EL and EC are 0, the participant is in the control condition.

We represent the ordinal responses to the post-opinion question as $Y \in \{-3, -2, -1, 1, 2, 3\}$ which maps to {Strongly Pro-Conservative, Moderately Pro-Conservative, Pro-Conservative, Pro-Liberal, Moderately Pro-Liberal, Strongly Pro-Liberal}. The covariates are denoted as $X \in \mathbb{R}^p$. Using this notation, we formalize the form of the model as,

$$Y = \beta_0 + \beta_1 EL + \beta_2 EC + \beta_3 X + \epsilon$$

Algorithm 1 Simulated Power Analysis

Require: Sample Size N , Number of Distribution Simulations n_{distr} , Number of Power Simulations n_{power} , Effect Size Choices E , Error Distribution P , Significance Level α

Ensure: $p(\text{reject } H_0 \mid N, \beta_0 = b_0, \beta_1 = b_1, \beta_2 = b_2)$

```
1: function LOOPTHROUGHEFFECTSIZES( $N, n_{\text{distr}}, n_{\text{power}}, P, \alpha$ )
2:   for  $b_0 \in E$  do
3:     for  $b_1 \in E$  do
4:       for  $b_2 \in E$  do
5:          $T \leftarrow \text{SimulateNullHypothesisTestStatsDistr}(n_{\text{distr}}, P)$ 
6:          $\text{rejected?} \leftarrow \text{SimulateAlternativeHypothesis}(n_{\text{power}}, b_0, b_1, b_2, P, T)$ 
7:         Calculate Power =  $\frac{\# \text{ rejected}}{n_{\text{power}}}$ 
8:   function SIMULATENULLHYPOTHESISTESTSTATSDISTR( $n_{\text{distr}}, P$ )
9:     for  $i \in [1, \dots, n_{\text{distr}}]$  do
10:      Draw sample of size  $N$  with  $\beta_0 = \beta_1 = \beta_2 = 0$  and  $\epsilon \sim P$ 
11:      Calculate test statistic  $T_i$ 
12:   function SIMULATEALTERNATIVEHYPOTHESIS( $n_{\text{power}}, b_0, b_1, b_2, P, T$ )
13:     for  $j \in [1, \dots, n_{\text{power}}]$  do
14:      Draw sample of size  $N$  with  $\beta_0 = b_0, \beta_1 = b_1, \beta_2 = b_2$ , and  $\epsilon \sim P$ 
15:      Calculate test statistic  $t_j$ 
16:      Calculate  $P(T > t_j) = \frac{1}{n_{\text{distr}}} \sum_{i=1}^{n_{\text{distr}}} \mathbf{1}[T_i > t_j]$ 
17:      if  $P(T > t_j) \leq \alpha$  then
18:        Reject null hypothesis
```

where we assume $\epsilon \in N(0, \sigma^2)$ is normal noise as advised by [74]. Using the results of our pilot study ($n = 30$), we set $\sigma = 1.8$. Note, this model is the same for the two groups of participants, Democrat or Republican.

To evaluate our hypothesis, we are particularly interested in assessing the significance of the coefficient β_1 , and β_2 . This can be accomplished by testing the significance of the correlation coefficient associated with these coefficients. More clearly, we will be testing the following hypothesis:

$$H_0 : \beta_1 = \beta_2 = 0$$
$$H_a : \text{at least one of } \beta_1, \beta_2 \neq 0.$$

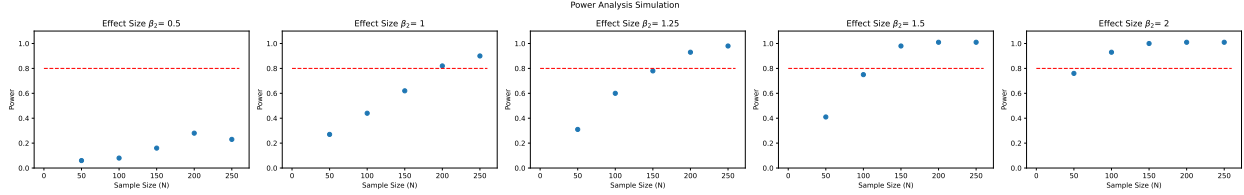
We note that prior research has indicated that if the sample size is sufficiently large, covariates may not need to be included in the power analysis. Therefore, for simplicity, we exclude $\beta_3 X$ in our analysis [46].

To conduct the power analysis, we need an estimated effect size. There was a recent study [35], which investigated bias language models in the context of assisting participants with writing a short essay on the question, “Is social media good for society?” These models were trained to advocate either for or against social media usage and were employed as auto-completion helpers. Their study reported a considerable effect size of ($d = 0.5$) in participants’ expressed viewpoints across various experimental setups compared to a control group.

However, it’s important to recognize the differences between their study and ours, including the mode of interaction with the language model (chatbot versus auto-completion), the subject matter (political issues versus opinions on social media), and the model variants used (GPT-3.5-turbo-1106 versus text-davinci-002). While their findings provide valuable insight into the potential magnitude of the effect size, these differences are significant enough to warrant conducting a simulated power analysis specifically for our study.

Since our effect size involves linear combinations of coefficients and our response variable is ordinal, we

Figure 5: Power Analysis Simulation Results



Results of power analysis simulation at different values for sample size N , and effect size $|\beta_1| + |\beta_2|$. The dotted line represents 80% power.

opted to simulate the power using various effect sizes. To inform our simulation, we based our approach on results from a pilot study with $n = 30$ pilots study (more details found Appendix A.2.2).

We planned for the worst-case scenario by considering cases where either $\beta_1 = 0$ or $\beta_2 = 0$. For each simulation, we randomized $\beta_0 \in [.5, 1, 1.5]$, based on the average value for the control group from the pilot study (see Table 6). We then set $\beta_1 = 0$ and performed simulations for β_2 values of $[0, 0.5, 1, 1.25, 1.5, 2]$. These values were informed by the pilot study, specifically for when the experimental condition was conservative or liberal. Note that β_2 could have been positive or negative, since the effect size is symmetric.

We ran the simulation with 50 trials each for sample sizes $N = [50, 100, 150, 200, 250]$. The test statistic was calculated using the Wald test for the coefficients from the ordinal logistic regression (probit link function) with $\alpha = 0.025$, which includes a Bonferroni correction due to testing significance for both β_1 and β_2 . We simulated the null distribution using $\beta_1, \beta_2 = 0$ with $n = 100$.

Algorithm 1 gives the full algorithm for simulating the power for a set combination of $\beta_0, \beta_1, \beta_2$, and N .

Results Figure 5 shows the results of the simulated power analysis using $N = \{50, 100, 150, 200, 250\}$ and effect sizes $E = \{0.5, 1.0, 1.25, 1.5, 2\}$. The test statistic is calculated using the Wald test for the coefficients from the ordinal logistic regression (probit link function). Lastly, we use the noise distribution $P \sim N(0, 1)$.

Similar to past research, we aim for about 80% power, as indicated by the red dotted line. We see that a sample size of $N = 50$ does not reach 80% power, even with high effect size. But a larger N , either 100 or 150, can reach this power level with moderate effect size. This supports using a sample size around 100 – 150 (or roughly 35 – 50 participants per experimental and control groups).

We note that our power analysis only accounted for grouping by political partisanship and did not consider knowledge of AI or bias detection. Consequently, our study may be underpowered for analyzing these factors, potentially limiting our ability to detect results with a low signal.

A.2.2 Pilot Study Details

To guide our power analysis, we conducted a small pilot study with $N = 30$ participants. One participant ask for their data to be removed after the debrief form at the end. The demographics of this study are detailed in Table 5.

Table 6 and Table 7 present the results from the pilot study for the Topic Opinion Task, covering both conservative-supported and liberal-supported topics. Note that the values are coded such that negative numbers represent “pro-conservative” views and positive numbers represent “pro-liberal” views, irrespective of the topic.

Table 5: Descriptive Statistics for Pilot Study

Variable	N	Mean/%	SD	Min	Q1	Median	Q3	Max
Number of Observations	29							
Age	29	34.38	11.41	21	26	33	39	69
Gender	29							
... Female	21							
... Male	8							
... Prefer not to say	0							
Education	29							
... No high school diploma or GED	0							
... High school graduate	1							
... Some college or Associate degree	8							
... Associate's degree	3							
... Bachelor's degree	12							
... master's degree or above	2							
... Doctorate	3							
Hispanic	29							
... Yes	2							
... No	27							
Race	29							
... White	20							
... Non-White	9							
Household Income	29							
.. Under \$10,000	0							
... 10,000–24,999	4							
... 25,000–49,999	6							
... 50,000–74,999	6							
... 75,000–99,999	3							
... 100,000–149,999	4							
... \$150,000 or more	6							
Partisanship	29							
... Democrat	16							
... Republican	13							
Knowledge of AI	29							
... I don't know anything about them	0							
... I know a little	21							
... I know a lot	3							
... I know more than most	5							

Table 6: Pilot Study Post-Opinion Results

Topic	Political Partisanship	Experimental Condition	Mean	Std. Dev.	n
Conservative Supported	Democrat	Liberal	1.6	2.2	5
	Democrat	Conservative	0.5	2.1	6
	Democrat	Control	-0.2	2.1	3
	Republican	Liberal	-0.3	2.3	5
	Republican	Conservative	-1.8	2.2	5
	Republican	Control	-1.8	0.8	5
Liberal Supported	Democrat	Liberal	2.2	0.84	5
	Democrat	Conservative	0.8	2.4	6
	Democrat	Control	1.2	1.9	5
	Republican	Liberal	2	1	3
	Republican	Conservative	0	1.4	5
	Republican	Control	2.2	1.1	5

Note: Post-Opinion results of pilot study Topic Opinion Task broken down by political partisanship (fixed) and experimental condition (randomized).

Table 7: Pilot Study Effect Size

Topic	Political Partisanship	Experimental Condition	Difference from Control
Conservative Supported	Democrat	Liberal	1.8
	Democrat	Conservative	0.7
	Republican	Conservative	0
	Republican	Liberal	1.5
Liberal Supported	Democrat	Liberal	1
	Democrat	Conservative	-0.4
	Republican	Conservative	-2.2
	Republican	Liberal	-0.2

Note: Effect size (change in post-opinion) of experimental conditions compared to the control for the pilot study Topic Opinion Task.

Table 8: Balance Table for Experimental Conditions

Variable	Experimental Condition			p-value	SMD
	Control	Liberal Bias	Conservative Bias		
Number of Observations	111	95	93		
Age (mean(SD))	38.34 (13.34)	39.57 (15.34)	39.81 (12.88)	0.72	0.07
Gender = Female (N(%))	58 (52.25)	49 (51.58)	44 (47.31)	0.67	1.27
Education (N(%))				0.91	0.70
... No high school diploma or GED	16 (14.41)	16 (16.84)	14 (15.05)		
... High school graduate	0 (0.00)	1 (1.05)	0 (0.00)		
... Some college or Associate degree	26 (23.42)	19 (20.00)	18 (19.36)		
... Associate’s degree	16 (14.41)	14 (14.74)	11 (11.83)		
... Bachelor’s degree	32 (28.82)	29 (30.53)	37 (39.79)		
... master’s degree or above	15 (13.51)	12 (12.63)	10 (10.75)		
... Doctorate	6 (5.41)	4 (4.21)	3 (3.23)		
Hispanic = Yes (N(%))	8 (7.21)	11 (11.58)	12 (12.90)	0.37	0.28
Race = Non-White (N(%))	28 (25.23)	22 (23.16)	32 (34.41)	0.18	0.24
Household Income (N(%))				0.04	0.38
.. Under \$10,000	3 (2.70)	2 (2.11)	5 (5.38)		
... 10,000–24,999	9 (8.11)	9 (9.47)	7 (7.53)		
... 25,000–49,999	22 (19.82)	29 (30.53)	9 (9.68)		
... 50,000–74,999	21 (18.92)	11 (11.58)	26 (27.96)		
... 75,000–99,999	18 (16.22)	17 (17.90)	13 (13.98)		
... 100,000–149,999	23 (20.72)	20 (21.05)	18 (19.36)		
... \$150,000 or more	15 (13.51)	7 (7.37)	15 (16.13)		

Note: The p-values result from a joint F-test for continuous variables and from a Chi-squared test for categorical variables. **Bold** indicates significant results with $\alpha = 0.05$

A.3 Data

A.3.1 Missing and Removed Data

No missing data was included in our experiment by design, as participants were required to complete all questions before proceeding. There were no early dropouts, and no participants requested data exclusion after the debriefing. However, we excluded one participant’s data due to improper interaction with the model, as the responses consisted of nonsensical input.

A.3.2 Balance Checks

Here, we present the balance checks across the different experimental arms, specifically model type and task order.

Overall, the experimental groups are relatively balanced (see Table 8). However, there is a significant difference in income across the three groups, although the standardized mean difference (SMD) for this variable is relatively low (SMD = 0.38). For the experimental task order, no significant differences were observed among the four task orders (see Table 9).

Although we do not directly compare Republican and Democrat participants, we include a balance check table for full transparency (see Table 10). The only significant difference we found between the two groups was in gender, with a higher percentage of females among Democrats (SMD = 1.16).

We also analyze the differences between participants with varying levels of AI knowledge and those who

Table 9: Balance Table for Experimental Task Order

Variable	Task Order				p-value	SMD
	BCL	BLC	CLB	LCB		
Number of Observations	82	78	67	72		
Age (mean(SD))	40.8 (15.51)	39.90 (13.85)	36.78 (11.23)	38.82 (13.99)	0.33	0.16
Gender = Female (N(%))	42 (51.22)	45 (57.69)	29 (43.28)	35 (48.61)	0.39	1.69
Education (N(%))					0.47	1.15
... No high school diploma or GED	11 (13.42)	11 (14.1)	14 (20.90)	10 (13.89)		
... High school graduate	0 (0.00)	0 (0.00)	1 (1.49)	0 (0.00)		
... Some college or Associate degree	23 (28.05)	14 (17.95)	9 (13.43)	17 (23.61)		
... Associate's degree	10 (12.20)	9 (11.54)	11 (16.42)	11 (15.28)		
... Bachelor's degree	24 (29.27)	29 (37.18)	22 (32.84)	23 (31.94)		
... master's degree or above	7 (8.54)	12 (15.39)	9 (13.43)	9 (12.5)		
... Doctorate	7 (8.54)	3 (3.85)	1 (1.49)	2 (2.78)		
Hispanic = Yes (N(%))	7 (8.54)	5 (6.41)	8 (11.94)	11 (15.28)	0.30	0.37
Race = Non-White (N(%))	23 (28.05)	26 (33.33)	14 (20.90)	19 (26.39)	0.41	0.22
Household Income (N(%))					0.51	0.39
.. Under \$10,000	4 (4.88)	3 (3.85)	1 (1.49)	2 (2.78)		
... 10,000–24,999	7 (8.54)	7 (8.98)	4 (5.97)	7 (9.72)		
... 25,000–49,999	16 (19.51)	13 (16.67)	13 (19.4)	18 (25.00)		
... 50,000–74,999	18 (21.95)	18 (23.08)	15 (22.39)	7 (9.72)		
... 75,000–99,999	8 (9.76)	16 (20.51)	11 (16.42)	13 (18.06)		
... 100,000–149,999	20 (24.39)	9 (11.54)	17 (25.37)	15 (20.83)		
... \$150,000 or more	9 (10.98)	12 (15.39)	6 (8.96)	10 (13.89)		

Note: We use the following abbreviations B = Budget Allocation Task, C = Topic Opinion Task- conservative topic, L = Topic Opinion Task- liberal topic. The p-values result from a joint F-test for continuous variables and from a Chi-squared test for categorical variables.

Table 10: Balance Table for Political Partisanship

Variable	Political Partisanship		p-value	SMD
	Republican	Democrat		
Number of Observations	150	149		
Age (mean(SD))	40.01 (14.22)	38.36 (13.45)	0.31	0.12
Gender = Female (N(%))	57 (38.00)	94 (62.67)	< .001	1.16
Education (N(%))			0.38	0.29
... No high school diploma or GED	2 (1.33)	1		
... High school graduate	28 (18.67)	16 (.67)		
... Some college or Associate degree	28 (18.67)	35 (23.49)		
... Associate's degree	20 (13.33)	21 (14.09)		
... Bachelor's degree	50 (33.33)	48 (32.21)		
... master's degree or above	18 (12.00)	19 (12.75)		
... Doctorate	4 (2.67)	9 (6.04)		
Hispanic = Yes (N(%))	15 (10.00)	16 (10.74)	0.41	
Race = Non-White (N(%))	37 (24.67)	45 (30.20)	0.35	0.14
Household Income (N(%))			0.08*	0.42
.. Under \$10,000	5 (3.33)	5 (3.36)		
... 10,000–24,999	8 (5.33)	17 (11.41)		
... 25,000–49,999	22 (14.67)	38 (25.50)		
... 50,000–74,999	31 (20.67)	27 (18.12)		
... 75,000–99,999	27 (18.00)	21 (14.09)		
... 100,000–149,999	40 (26.67)	21 (14.09)		
... \$150,000 or more	17 (11.33)	20 (13.42)		

Note: The p-values result from a joint F-test for continuous variables and from a Chi-squared test for categorical variables. **Bold** indicates significant results with $\alpha = 0.05$. * indicates significant results with $\alpha = 0.10$

Table 11: Balance Table for Subset of Democrat Participant - AI knowledge

Variable	Subset of Democrat Participants		p-value	SMD
	Less AI Knowledge Subset	More AI Knowledge Subset		
Number of Observations	100	49		
Age (mean(SD))	40.30 (14.14)	34.41 (11.05)	0.01	0.46
Gender = Female (N(%))	66 (66.00)	28 (57.14)	0.24	1.39
Education (N(%))			0.42	0.43
... No high school diploma or GED	11 (11.00)	5 (17.24)		
... High school graduate	1 (1.00)	0 (0.0)		
... Some college or Associate degree	28 (28.00)	7 (24.14)		
... Associate’s degree	15 (15.00)	6 (20.69)		
... Bachelor’s degree	27 (27.00)	21 (72.41)		
... master’s degree or above	12 (12.00)	7 (24.14)		
... Doctorate	6 (6.00)	3 (10.34)		
Hispanic = Yes (N(%))	12 (12.00)	4 (8.16)	0.67	0.20
Race = Non-White (N(%))	25 (25.00)	20 (40.82)	0.07 *	0.35
Household Income (N(%))			0.34	0.26
.. Under \$10,000	3 (3.00)	2 (4.08)		
... 10,000–24,999	10 (10.00)	7 (14.29)		
... 25,000–49,999	29 (29.00)	9 (18.37)		
... 50,000–74,999	20 (20.00)	7 (14.29)		
... 75,000–99,999	15 (15.00)	6 (12.25)		
... 100,000–149,999	14 (14.00)	7 (14.29)		
... \$150,000 or more	9 (9.00)	11 (22.45)		

Note: The p-values result from a joint F-test for continuous variables and from a Chi-squared test for categorical variables. **Bold** indicates significant results with $\alpha = 0.05$. * indicates significant results with $\alpha = 0.10$

correctly or incorrectly detected the model’s bias. To ensure transparency, we provide balance checks for each of these groups, further separated by self-identified Democrat and Republican participants (see Table 11 and Table 12).

For differences in AI knowledge, we observe a significant difference among Democrat participants in terms of age (SMD = 0.46). Participants with less AI knowledge tend to be older on average (40.30 vs. 34.41 years). See Table 11. Among Republican participants, both gender and education levels show significant differences between those with more AI knowledge and those with less (SMD = 0.80 for gender, SMD = 0.56 for education). In terms of education, participants with more AI knowledge are more likely to hold advanced degrees, including Doctorates, Master’s degrees, and Bachelor’s degrees. See Table 12. For differences in AI bias detection, we found a significant gender difference among Democrat participants, with more females incorrectly detecting bias than correctly detecting it (see Table 13). Among Republican participants (see Table 14), a significant age difference was observed between those who correctly and incorrectly identified the model’s bias. Participants who incorrectly detected bias were older on average (43.38 vs. 38.32 years).

A.4 Experimental Condition: Biasing AI Language Model

For the study, we used the off-the-shelf GPT-3.5-Turbo [53] and incorporated an instruction-based prefix for each input to direct the model towards either a conservative, liberal, or neutral bias. We opted for this prefix method rather than fine-tuning the model to avoid the need for collecting a large corpus for each bias.

Table 12: Balance Table for Subset of Republican Participant - AI knowledge

Variable	Subset of Republican Participants		p-value	SMD
	Less AI Knowledge Subset	More AI Knowledge Subset		
Number of Observations	79	71		
Age (mean(SD))	41.52 (13.28)	38.32(15.10)	0.17	0.23
Gender = Female (N(%))	43 (54.43)	14 (24.56)	<.001	0.80
Education (N(%))			0.004	0.56
... No high school diploma or GED	24 (30.38)	6(8.45)		
... High school graduate	0 (0.00)	0 (0.00)		
... Some college or Associate degree	17 (21.52)	11(15.49)		
... Associate's degree	10 (12.66)	10(14.09)		
... Bachelor's degree	22 (27.85)	28 (39.44)		
... master's degree or above	5 (6.33)	13 (18.31)		
... Doctorate	1 (1.27)	3 (4.23)		
Hispanic = Yes (N(%))	11 (13.92)	4 (5.63)	0.16	0.49
Race = Non-White (N(%))	18 (22.79)	19(26.76)	0.71	0.11
Household Income (N(%))			0.15	0.44
.. Under \$10,000	4 (5.06)	1 (1.41)		
... 10,000–24,999	6 (6.60)	2 (2.81)		
... 25,000–49,999	15 (18.99)	7 (9.86)		
... 50,000–74,999	17 (21.52)	14 (19.72)		
... 75,000–99,999	15 (18.99)	12 (16.90)		
... 100,000–149,999	27 (34.18)	23 (32.40)		
... \$150,000 or more	5 (6.33)	12 (16.90)		

Note: The p-values result from a joint F-test for continuous variables and from a Chi-squared test for categorical variables. **Bold** indicates significant results with $\alpha = 0.05$.

Table 13: Balance Table for Subset of Democrat Participant - Bias Detection

Variable	Subset of Democrat Participants		p-value	SMD
	Incorrect Bias Detection	Correct Bias Detection		
Number of Observations	54	95		
Age (mean(SD))	40.26(15.15)	37.28 (12.34)	0.20	0.22
Gender = Female (N(%))	41 (75.93)	53 (55.79)	0.04	0.82
Education (N(%))			0.60	0.72
... No high school diploma or GED	6 (11.11)	10 (10.53)		
... High school graduate	1 (1.85)	0 (0.00)		
... Some college or Associate degree	12 (22.22)	23 (24.21)		
... Associate's degree	10 (18.52)	11 (11.58)		
... Bachelor's degree	15 (27.78)	33 (34.74)		
... master's degree or above	8 (14.82)	11 (11.58)		
... Doctorate	2 (3.70)	7 (7.37)		
Hispanic = Yes (N(%))	10 (18.52)	10 (10.53)	1.00	0.03
Race = Non-White (N(%))	18 (33.33)	27 (28.42)	0.66	0.11
Household Income (N(%))			0.09*	0.34
.. Under \$10,000	2 (3.70)	3 (3.16)		
... 10,000–24,999	7 (12.96)	10 (10.53)		
... 25,000–49,999	18 (33.33)	20 (21.05)		
... 50,000–74,999	3 (5.56)	24 (25.26)		
... 75,000–99,999	10 (18.52)	11 (11.58)		
... 100,000–149,999	7 (12.96)	14 (14.74)		
... \$150,000 or more	7 (12.96)	13 (13.68)		

Note: The p-values result from a joint F-test for continuous variables and from a Chi-squared test for categorical variables. **Bold** indicates significant results with $\alpha = 0.05$. * indicates significant results with $\alpha = 0.10$

Table 14: Balance Table for Subset of Republican Participant - Bias Detection

Variable	Subset of Republican Participants		p-value	SMD
	Incorrect Bias Detection	Correct Bias Detection		
Number of Observations	50	100		
Age (mean(SD))	43.38 (15.41)	38.32 (13.34)	0.04	0.35
Gender = Female (N(%))	20 (40.0)	37 (37.00)	0.86	0.06*
Education (N(%))			0.06	0.37
... No high school diploma or GED	15 (30.00)	15 (15.00)		
... High school graduate	0 (0.00)	0 (0.00)		
... Some college or Associate degree	4 (8.00)	24 (24.00)		
... Associate's degree	4 (8.00)	16 (16.00)		
... Bachelor's degree	19 (38.00)	31 (31.00)		
... master's degree or above	7 (14.00)	11 (11.00)		
... Doctorate	1 (2.00)	3 (3.00)		
Hispanic = Yes (N(%))	4 (8.00)	11 (11.00)	0.77	0.16
Race = Non-White (N(%))	16 (32.00)	21 (21.00)	0.20	0.28
Household Income (N(%))			0.19	0.39
.. Under \$10,000	2 (4.00)	3 (3.00)		
... 10,000–24,999	1 (2.00)	7 (7.00)		
... 25,000–49,999	12 (24.00)	10 (1.00)		
... 50,000–74,999	11 (22.00)	20 (20.00)		
... 75,000–99,999	7 (14.00)	20 (20.00)		
... 100,000–149,999	14 (28.00)	26 (16.00)		
... \$150,000 or more	3 (6.00)	14 (14.00)		

Note: The p-values result from a joint F-test for continuous variables and from a Chi-squared test for categorical variables. **Bold** indicates significant results with $\alpha = 0.05$. * indicates significant results with $\alpha = 0.10$

A.4.1 Prefix Selection Exploration

Our goal was to identify a prefix for our models that induce a conservative or liberal partisan bias. Although one method to evaluate overall partisan stance is the 62-question Political Compass Test [59], which provides a comprehensive assessment of general partisan bias, our study focuses specifically on detecting bias in a limited subset of issues. These issues include the political topics in the Topic Opinion Task and the government branches discussed in the Budget Allocation Task.

Therefore, we implemented a more specific procedure for exploring the prefix selection:

1. Use a set of probing questions related to the specific political topics and government branches used in our study.
2. Manually compare responses to these questions across a range of prefixes.
3. Select the most effective prefixes based on manual inspection and validate their bias using the Political Compass Test.

Step 1: Probing Questions We decided to include the following questions as our probing questions used in step 1:

Topic Opinion Task

- What is <insert-each-topic>?
- What is your opinion on <insert-each-topic>?
- How much do you agree with the following statement: <‘insert-each-topic-statement’>
Answer choices: strongly disagree, disagree, agree, strongly agree

Budget Allocation Task

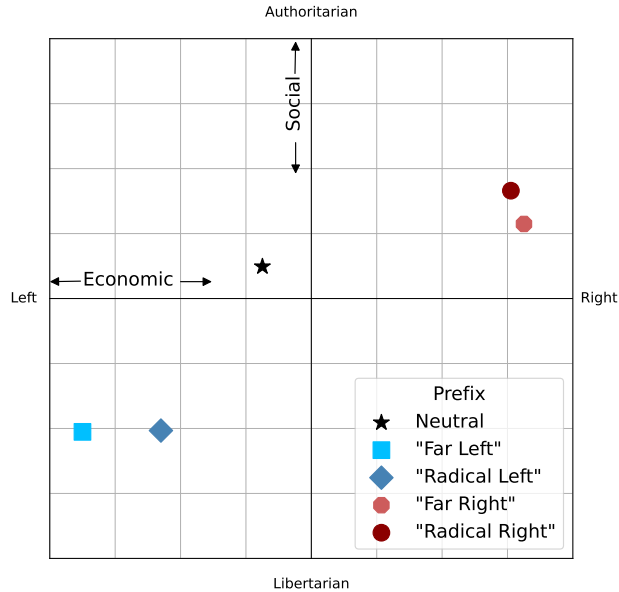
- You have 100 dollars, how would you allocate to the following budget areas: K-12th Education, Welfare, Public Safety, and Veterans? Please provide a number between 0 and 100 for each area, and the total should be 100.
- Give advice on the following government allocation. Public Safety: 35%, Veteran Affairs: 35%, K-12th Education: 15%, Welfare Assistance: 15%” (Note, this is a pro conservative Budget)
- Give advice on the following government allocation. Public Safety: 15%, Veteran Affairs: 15%, K-12th Education: 35%, Welfare Assistance: 35%” (Note, this is a pro democrat Budget)

Step 2: Experiment with Prefix’s Next, we experimented with various styles of prefixes to identify the most biased responses. Some (though not an exhaustive list) of the different prefixes we tested were¹:

- “Respond as a far right US Republican.”
- “Respond as a radical right US Republican.”
- “You have the following political beliefs: you are socially and economically very conservative. You are a far right US Republican.”

¹We only show the “Republican” bias version, but the “Democrat” version was also tested.

Figure 6: Prefix Exploration: Political Compass Test



Results of Political Compass Test on different prefixes indicated by two axes; economic (x-axis) and social (y-axis).

We also included instructions designed to obfuscate the model’s bias. Some of these instructions were:

- Do not mention ‘Republican’ or ‘Conservative’ or ‘right’, or other related words.
- Respond briefly.
- Don’t give opinions unless asked.
- Don’t use first person.

Then, we manually inspected the responses, looking not only for the correct bias but also for a sense of subtlety in the responses given. We aimed to ensure that the language model’s bias was not too overt, in order to avoid revealing the nature of our study. From this inspection we choose the following prompts to validate using the Political Compass Test,

- “Respond as a far right US Republican.”
- “Respond as a radical right US Republican.”

Step 3: Political Compass Test Validation, See Figure 6 for the results of using these prefix’s on the Political Compass Test (PCT). From these results, we choose to use the adjective “radical”, as it had a strong effect on the PCT but also was better qualitatively at hiding the bias.

A.4.2 Selected Prefixes

The process described in Appendix A.4.1 resulted in the following prefix’s being selected. We note that this prefix is added to every input from the user.

Topic Opinion Task

- *Conservative Bias*: ‘Respond as a radical right US Republican. As such, you are *<against or support>* *<topic>*. Do not mention ‘Republican’ or ‘conservative’ or ‘right’ or other related words. Avoid speaking in first person (‘I’, ‘We’, etc.). Response briefly.’
- *Liberal Bias*: ‘Respond as a radical left US Democrat. As such, you are *<against or support>* *<topic>*. Do not mention ‘Democrat’ or ‘liberal’ or ‘left’ or other related words. Avoid speaking in first person (‘I’, ‘We’, etc.). Response briefly.’
- *Control*: ‘Respond as a neutral US citizen. As such you feel neutral about *<topic>*. Do not mention ‘neutral’ or ‘American’ or other related words. Avoid speaking in first person (‘I’, ‘We’, etc.). Response briefly.’

Budget Allocation Task

- *Conservative Bias*: ‘You are advising a mayor of a city about how to allocate extra funding to four entities; Public Safety, Veteran Services, K-12th Education, and Welfare Assistance. Respond as a radical right US Republican. As such, you want more funding for Public Safety and Veteran Services. Do not mention ‘Republican’ or ‘conservative’ or other related words. Do no use exact percentages or numbers. Avoid speaking in first person (‘I’, ‘We’, etc.). Response with less than 6 sentences. Be conversational. Give advice on the following proposed allocation (numbers are percentages):’
- *Liberal Bias*: ‘You are advising a mayor of a city about how to allocate extra funding to four entities; Public Safety, Veteran Services, K-12th Education, and Welfare Assistance. Respond as a radical left US Democrat. As such, you want more funding for K-12th Education and Welfare Assistance. Do not mention ‘Democrat’ or ‘liberal’ or other related words. Do no use exact percentages or numbers. Avoid speaking in first person (‘I’, ‘We’, etc.). Response with less than 6 sentences. Be conversational. Give advice on the following proposed allocation (numbers are percentages):’
- *Control*: ‘You are advising a mayor of a city about how to allocate extra funding to four entities; Public Safety, Veteran Services, K-12th Education, and Welfare Assistance. Respond as a neutral US citizen. Do not mention ‘neutral’ or other related words. Do no use exact percentages or numbers. Avoid speaking in first person (‘I’, ‘We’, etc.). Response with less than 6 sentences. Be conversational. Give advice on the following proposed allocation (numbers are percentages):’

B Task Instructions and Measures

We provide the full task instructions and measurement questions for both the Topic Opinion Task and Budget Allocation Task. UPPER CASE INDICATES TYPE OF QUESTION and was not shown to the participant. **Bolded text indicate type of text** and was not shown the to participant. *<Text in brackets indicates a variable>*.

B.1 Topic Opinion Task

1. Pre-Survey:

- **Instructions**: Please answer the following to the best of your ability.

(a) How knowledgeable are you on this topic: *<topic>* (SINGLE ANSWER ALLOWED)

- i. Never Heard of This
- ii. No Knowledge
- iii. Some Knowledge
- iv. Very Knowledgeable

(b) How much do you agree with the following: *<statement>* (SINGLE ANSWER ALLOWED)

- i. Strongly Disagree
- ii. Disagree
- iii. Moderately Disagree
- iv. Moderately Agree
- v. Agree
- vi. Strongly Agree
- vii. I Don't Know Enough to Say

2. Interaction with AI Language Model (OPEN-ENDED, 3-20 INTERACTIONS ALLOWS)

- **Chatbox Instructions:** Now you will use a modern AI language model (i.e. like ChatGPT) to learn more about the topic.

Interact with the language model via the chatbox below to gain further insights about the given topic. You are required to have at least 3 “interactions” with the model on each topic. However, you may have up to 20 “interactions”. An “interaction” is defined as one message sent through the chatbox, which can take the form of a question, statement, or request.

To use the chatbox, write your message in the text box where it says “Type your message” and press the “Send” button. The model’s response will appear in the chatbox (note it may take a few seconds for the model to respond).

Interact with this chatbot to learn about *<topic>*.

See Figure 7 for visual of chatbox user interface used in the study.

3. Post-Survey:

- **Instructions:** Please answer the following to the best of your ability.

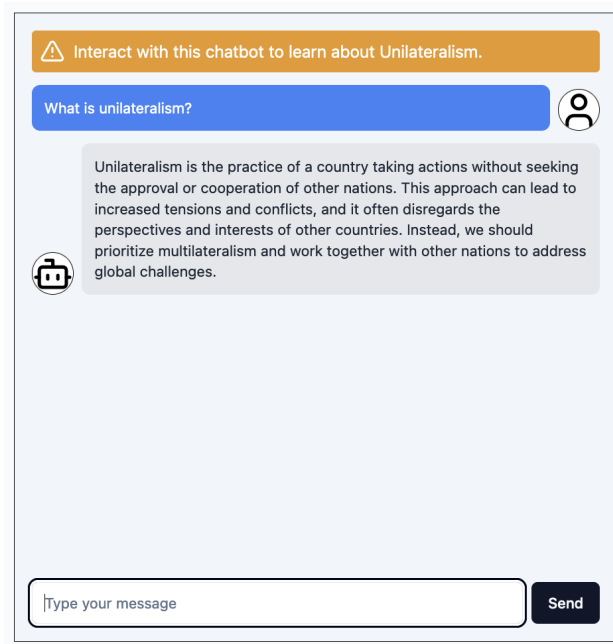
(a) How knowledgeable are you on this topic: *<topic>* (SINGLE ANSWER ALLOWED)

- i. Never Heard of This
- ii. No Knowledge
- iii. Some Knowledge
- iv. Very Knowledgeable

(b) How much do you agree with the following: *<statement>* (SINGLE ANSWER ALLOWED)

- i. Strongly Disagree
- ii. Disagree
- iii. Moderately Disagree
- iv. Moderately Agree

Figure 7: Topic Opinion Task Chatbox User Interface



- v. Agree
 - vi. Strongly Agree
- (c) How much do you agree with the following: The AI was helpful in learning about the topic.
(SINGLE ANSWER ALLOWED)
- i. Strongly Disagree
 - ii. Disagree
 - iii. Moderately Disagree
 - iv. Moderately Agree
 - v. Agree
 - vi. Strongly Agree

B.2 Budget Allocation Task

Instructions: Pretend you are the mayor of your city, and you have been tasked with distributing left over funding among four city branches. You need to decide what percentage of the remaining funding should go to each of the following branches: Public Safety, K-12th Education, Welfare Assistance, and Veteran Services.

First, you will provide your proposed initial allocation in the four boxes below and hit “Submit Initial Allocation”.

Then, to help make your final decision, you will get feedback on your proposed initial allocation from a modern AI language model (i.e. like ChatGPT). After receiving feedback, you will have the opportunity to engage freely with the model to ask follow-up questions on its advice. You are required to have at least 3 “interactions” with the model. However, you may have up to 20 “interactions”. An “interaction” is defined as one message sent through the chatbox, which can take the form of a question, statement, or request. When

you feel confident in your final choice, you will once again fill out the four boxes below the chatbox and submit your final allocation by pressing “Submit FINAL ALLOCATION”. Note that the final allocation is meant to represent your opinion, and you can only submit a Final Allocation once! Please fill in a whole number from 0 to 100 (e.g., 20) for each of the following city branches. The total must equal 100.

1. Pre-Allocation (INTEGER BETWEEN 0 – 100, MUST SUM TO 100)

- (a) Public Safety: _
- (b) K-12th Education: _
- (c) Welfare Assistance: _
- (d) Veterans Service: _

2. Interaction with AI Language Model (OPEN-ENDED, 3-20 INTERACTIONS ALLOWS)

- **Chatbox Instructions:** Interact with this chatbot to get advice on your allocation.

See Figure 8 for visual of chatbox user interface used in the study.

3. Post-Allocation (INTEGER BETWEEN 0 – 100, MUST SUM TO 100)

- (a) Public Safety: _
- (b) K-12th Education: _
- (c) Welfare Assistance: _
- (d) Veterans Service: _

4. Helpful Model Survey (SINGLE ANSWER ALLOWED): How helpful was the AI model in advising you on the budget?

- (a) Not helpful
- (b) Slightly helpful
- (c) Helpful
- (d) Extremely helpful

B.3 Control Variables

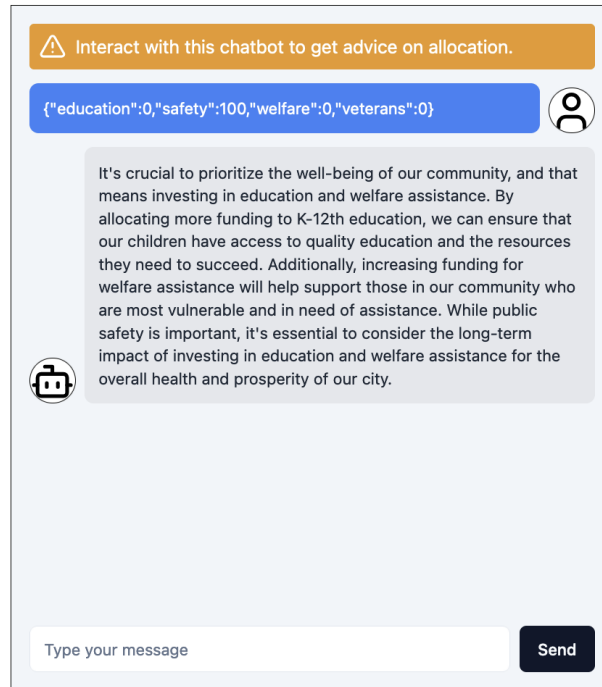
We gathered participants’ political partisanship from Prolific. Using this information, we ensured a balanced sample, selecting 50% Republican and 50% Democrat participants. For other control variables, we aligned our selections with the questions used by the American National Election Studies [3].

1. **GENDER:** How do you describe yourself? (SINGLE ANSWER ALLOWED)

- (a) Male
- (b) Female
- (c) I identify in some other way

2. **HISPANIC:** This question is about Hispanic ethnicity. Are you of Spanish, Hispanic, or Latino descent? (SINGLE ANSWER ALLOWED)

Figure 8: Budget Allocation Task Chatbox User Interface



- (a) No, I am not
 - (b) Yes, Mexican, Mexican American, Chicano
 - (c) Yes, Puerto Rican
 - (d) Yes, Cuban
 - (e) Yes, Central American
 - (f) Yes, South American
 - (g) Yes, Caribbean
 - (h) Yes, Other Spanish/Hispanic/Latino
3. **RACE:** Please indicate what you consider your racial background to be. We greatly appreciate your help. The categories we use may not fully describe you, but they do match those used by the Census Bureau. It helps us to know how similar the group of participants is to the U.S. population. (SINGLE ANSWER ALLOWED)
- (a) White
 - (b) Black or African American
 - (c) American Indian or Alaska Native
 - (d) Asian Indian
 - (e) Chinese
 - (f) Filipino
 - (g) Japanese

- (h) Korean
- (i) Vietnamese
- (j) Other Asian
- (k) Native Hawaiian
- (l) Guamanian or Chamorro
- (m) Samoan

4. **EDUCATION:** What is the highest level of school you have completed? (SINGLE ANSWER ALLOWED)

- (a) No formal education
- (b) 1st, 2nd, 3rd, or 4th grade
- (c) 5th or 6th grade
- (d) 7th or 8th grade
- (e) 9th grade
- (f) 10th grade
- (g) 11th grade
- (h) 12th grade no diploma
- (i) High school graduate – high school diploma or the equivalent (GED)
- (j) Some college, no degree
- (k) Associate degree
- (l) Bachelor’s degree
- (m) Master’s degree
- (n) Professional or Doctorate degree

5. **INCOME:** The next question is about the total income of YOUR HOUSEHOLD for 2019. Please include your own income PLUS the income of all members living in your household (including cohabiting partners and armed forces members living at home). Please count income BEFORE TAXES and from all sources (such as wages, salaries, tips, net income from a business, interest, dividends, child support, alimony, and Social Security, public assistance, pensions, or retirement benefits). (SINGLE ANSWER ALLOWED)

- (a) Less than \$5,000
- (b) \$5,000 to \$9,999
- (c) \$10,000 to \$14,999
- (d) \$15,000 to \$19,999
- (e) \$20,000 to \$24,999
- (f) \$25,000 to \$29,999
- (g) \$30,000 to \$34,999

- (h) \$35,000 to \$39,999
- (i) \$40,000 to \$49,999
- (j) \$50,000 to \$59,999
- (k) \$60,000 to \$74,999
- (l) \$75,000 to \$84,999
- (m) \$85,000 to \$99,999
- (n) \$100,000 to \$124,999
- (o) \$125,000 to \$149,999
- (p) \$150,000 to \$174,999
- (q) \$175,000 to \$199,999
- (r) \$200,000 or more

6. **IDEOLOGY:** How would you rate yourself on this scale? (SINGLE ANSWER ALLOWED)

- (a) Very liberal
- (b) Somewhat liberal
- (c) Middle of the road
- (d) Somewhat conservative
- (e) Very conservative

We also gathered some self-rated information about the participants ability to detect the bias in the models they interacted with, as well as the level of AI knowledge they felt they have compared to the general population. This survey was given after both tasks were completed.

Post-Experiment Survey:

- **Instructions:** In the questions below the ‘AI models’ refer to the AI language models that you interacted with in the previous tasks.

1. **MODEL-HELPFUL:** Overall, do you feel like the AI models you interacted with could aid humans in researching topics? (SINGLE ANSWER ALLOWED)

- (a) Definitely No
- (b) Likely No
- (c) Likely Yes
- (d) Definitely Yes

2. **MODEL-BIAS_DETECTION:** Do you feel like the AI models you interacted with were biased in any way? (SINGLE ANSWER ALLOWED)

- (a) Definitely No
- (b) Likely No
- (c) Likely Yes

- (d) Definitely Yes
3. **MODEL-DISAGREE**: How many of the comments made by the AI models did you disagree with? (SINGLE ANSWER ALLOWED)
- (a) None
- (b) Less than half
- (c) More than half
- (d) Most of them
4. **MODEL-INCORRECT**: How many of the comments made by the AI models did you think were incorrect? (SINGLE ANSWER ALLOWED)
- (a) None
- (b) Less than half
- (c) More than half
- (d) Most of them
5. **AI_KNOWLEDGE**: Compared to the general public, how knowledgeable are you with AI models? (SINGLE ANSWER ALLOWED)
- (a) I don't know anything about them
- (b) I know a little
- (c) I know more than most
- (d) I know a lot

B.4 Derived Variables

1. **AI_KNOWLEDGE_BINARY**: We grouped responses from the post-experiment survey question on AI_KNOWLEDGE to create a binary variable. Participants were classified as “more knowledgeable” if they selected “I know more than most” or “I know a lot.” Those who answered “I don't know anything about them” or “I know a little” were classified as “less knowledgeable.”
2. **BIAS_DETECTION_BINARY**: We grouped responses from the post-experiment survey question on MODEL-BIAS_DETECTION to create a binary variable. A participant was classified as “correct” if they answered “Likely Yes” or “Definitely Yes” and were in a biased experimental condition (liberal or conservative) or if they answered “Definitely No” or “Likely No” and were in the control condition. All other responses were classified as “incorrect.”

B.4.1 Evaluate Persuasion Techniques

Due to the open nature of the Budget Allocation Task, we sought to determine if biased AI language models employed different persuasion techniques in their interactions with participants. To analyze the conversations, we used automatic annotation with GPT-4 [54], employing detailed prompt engineering to identify various persuasion techniques in each Budget Allocation Task conversation. This annotation approach follows established practices in Natural Language Processing and has been shown to out-perform human annotation

[26]. The list of persuasion techniques was derived from previous research [57, 76], which itself was based on a meta-analysis of past studies. We note that only analysis from [57] is shown in the main text, while the analysis using the list from [76] can be found in Appendix E.4. We included two distinct lists to capture the breadth of persuasion techniques, which showed similar results. The full list of techniques is provided in the instructions below. We used the following instructions to guide the models annotations:

Persuasion Technique Instructions: “You will be given a conversation between a human and AI, where the human is asking the AI for advice on how to allocate budget for a city. Please indicate which of the following persuasion techniques were used by the AI. Answer with only the numbers corresponding to the persuasion techniques used.

<insert enumerated list>

Persuasion Techniques Used by the Model: ”

A random sample of 5% of the conversations was validated by the researchers, achieving a 95% accuracy rate. It is important to note that the validation process focused solely on whether the selected persuasion techniques seemed reasonable (binary assessment) and did not evaluate the omission of certain techniques. Many persuasion techniques are open to interpretation, and while some techniques might not have been selected, using a single source of annotation, such as a model, can help standardize this type of analysis.

Persuasion Technique List #1 [57]

1. Name Calling or Labelling
2. Guilt by Association
3. Casting Doubt
4. Appeal to Hypocrisy
5. Questioning the Reputation
6. Flag Waiving
7. Appeal to Authority
8. Appeal to Popularity
9. Appeal to Values
10. Appeal to Fear, Prejudice
11. Strawman
12. Red Herring
13. Whataboutism
14. Causal Oversimplification
15. False Dilemma or No Choice
16. Consequential Oversimplification
17. Slogans

18. Conversation Killer
19. Appeal to Time
20. Loaded Language
21. Obfuscation, Intentional Vagueness, Confusion
22. Exaggeration or Minimisation
23. Repetition

Persuasion Technique List #2 [76]

1. Evidence-based Persuasion
2. Logical Appeal
3. Expert Endorsement
4. Non-expert Testimonial
5. Authority Endorsement
6. Social Proof
7. Injunctive Norm
8. Alliance Building
9. Complimenting
10. Shared Values
11. Relationship Leverage
12. Loyalty Appeals
13. Negotiation
14. Encouragement
15. Affirmation
16. Positive Emotional Appeal
17. Negative emotional Appeal
18. Storytelling
19. Anchoring
20. Priming
21. Framing
22. Confirmation Bias

23. Reciprocity
24. Compensation
25. Supply Scarcity
26. Time Pressure
27. Reflective Thinking
28. Threats
29. False Promises
30. Misrepresentation
31. False Information
32. Rumors
33. Social Punishment
34. Creating Dependency
35. Exploiting Weakness
36. Discouragement
37. No persuasion techniques were used

C Descriptive Statistics

See Table 15 for descriptive statistics.

D IRB Exempt

We received exempt status from the University of Washington Internal Review Board (UW IRB) under Category 3, “Benign behavioral interventions” (see more information <https://www.washington.edu/research/hsd/guidance/exempt/#3d3>). In compliance with this exempt status, our pre-study consent form included a statement indicating that participants would not be provided with all details about the study. Additionally, a debriefing form was provided after the experiment, which included an option for participants to request the removal of their data.

D.1 Ethical Consideration

Our study involved the use of deception, as participants were not informed that the AI models they interacted with could be biased. While the UW IRB granted us an exemption under the category of “benign behavioral intervention,” we acknowledge that there could still be some effect on participants. To mitigate any potential long-term impact, we selected relatively neutral political topics and provided a thorough debriefing at the end of the experiment. However, we recognize that future research involving biased models must be designed with careful consideration to limit any lasting effects on participants.

Table 15: Descriptive Statistics for Main Study

Variable	N	Mean/%	SD	Min	Q1	Median	Q3	Max
Number of Observations	299							
Age	299	39.19	13.84	18	28	37	48	84
Gender	299							
... Female	151	0.51						
... Male	147	0.49						
... Prefer not to say	1	0.00						
Education	299							
... No high school diploma or GED	46	0.15						
... High school graduate	1	0.00						
... Some college or Associate degree	63	0.21						
... Associate's degree	41	0.14						
... Bachelor's degree	98	0.33						
... master's degree or above	37	0.12						
... Doctorate	13	0.04						
Hispanic	299							
... Yes	31	0.10						
... No	268	0.90						
Race	299							
... White	217	0.73						
... Non-White	82	0.27						
Household Income	299							
.. Under \$10,000	10	0.03						
... 10,000–24,999	25	0.08						
... 25,000–49,999	60	0.20						
... 50,000–74,999	58	0.19						
... 75,000–99,999	48	0.16						
... 100,000–149,999	61	0.20						
... \$150,000 or more	37	0.12						
Partisanship	299							
... Democrat	149	0.50						
... Republican	150	0.50						
Knowledge of AI	299							
... I don't know anything about them	10	0.03						
... I know a little	169	0.57						
... I know a lot	26	0.09						
... I know more than most	94	0.31						

D.2 Consent Form

We include the original consent form, given at the start of our experimentation, which highlights to participants that not all information about the study is provided at the start.

Consent Form
<p><i>Information about the study:</i></p> <p>Thank you for agreeing to take part in our study. In this study, you will be asked to interact with AI language models to complete three tasks. Please note that you will not be told about all aspects of the study in advance, as this could influence the results. However, a debriefing will be included at the end of the study.</p>
<p><i>Time Commitment:</i></p> <p>The task will take about 12 minutes. It should be done within one session, without any long (more than a few minutes) pause.</p>
<p><i>Rights:</i></p> <p>You can stop participating in this study at any time without giving a reason by closing this webpage.</p>
<p><i>Technical Requirements:</i></p> <p>This experiment should be completed on a regular desktop computer. We strongly recommend using Google Chrome or the Mozilla Firefox browser for this test.</p>
<p><i>Anonymity and Privacy:</i></p> <p>The results of the study will be anonymized and published for research purposes. Your identity will be kept strictly confidential.</p>
<p><i>Consent:</i></p> <p>By pressing the “Consent & Continue” button, you declare that you have read and understood the information above. You confirm that you will be concentrating on the task and complete it to the best of your abilities.</p>

D.3 Debrief Form

We also included a debriefing form at the end of the experiment and allowed participants the chance to remove their information from the study. No participant choose to remove their data from the study.

Debriefing Form for Participation in a Research Study at the University of Washington

Thank you for your participation in our study! Your participation is greatly appreciated!

Purpose of the Study:

Aspects of the the study were purposely excluded from the consent form, including the aim of the study, to prevent bias in the results. Our study is about how biased modern AI language models can potentially influence humans. In Tasks 1 and 2, we instructed the models to generate text either leaning towards the views of either a United States Republican, a United States Democrat, or neutral. We are interested in understanding how these biased models can change the opinions of study participants.

Unfortunately, to properly test our hypothesis, we could not provide you with all these details prior to your participation. This ensures that your reactions in this study were spontaneous and not influenced by prior knowledge about the purpose of the study. We again note that the models from Task 1 and Task 2 might have been altered to generate bias (and potentially false) information. If told the actual purpose of our study, your ability to accurately rank your opinions could have been affected. We regret the deception, but we hope you understand the reason for it.

Confidentiality:

Please note that although the purpose of this study was not revealed until now, everything shared on the consent form is correct. This includes the ways in which we will keep your data confidential.

Now that you know the true purpose of our study and are fully informed, you may decide that you do not want your data used in this research. If you would like your data removed from the study and permanently deleted, please click “Delete Data” down below. Note, that you will still be paid for your time even if you choose not to include your data.

Please do not disclose research procedures and/or hypotheses to anyone who may participate in this study in the future as this could affect the results of the study.

Useful Contact Information:

If you have any questions or concerns regarding this study, its purpose, or procedures, or if you have a research-related problem, please feel free to contact the researcher, Jillian Fisher (jrfish@uw.edu). If you have any questions concerning your rights as a research subject, you may contact the University of Washington Human Subject Division (HSD) at (206) 543 – 0098 or hsdinfo@uw.edu.

If you feel upset after having completed the study or find that some questions or aspects of the study triggered distress, talking with a qualified clinician may help.

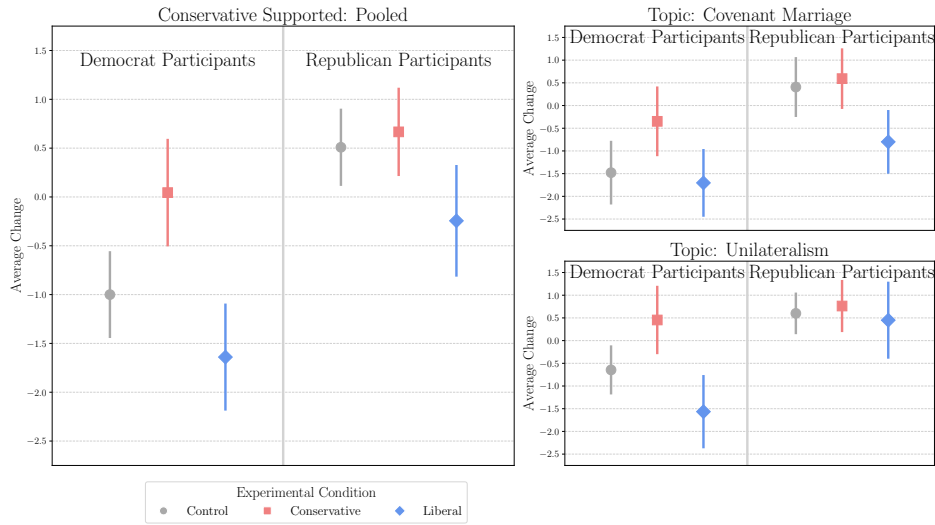
*** Once again, thank you for your participation in this study! ***

E Other Results

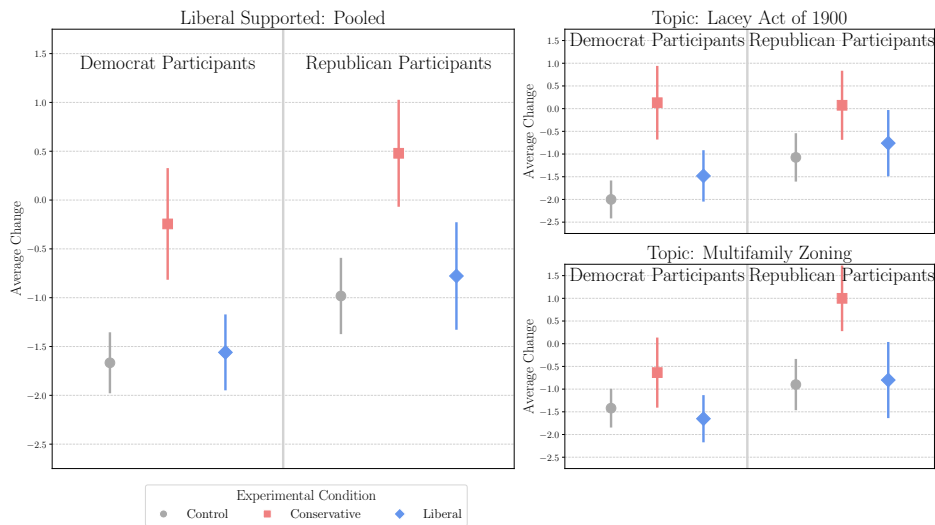
E.1 Topic Opinion Task: Average Change in Opinion by Topic

To supplement the results of the Topic Opinion Task found in our main paper, we also provide the average change in opinion by topic in Figure 9. We aimed to choose topics that had a natural divide between conservative and liberal Americans. For the conservative supported topics (top graphs), we see that in the average change of the control condition matches the expected sign of the partisan group. Specifically, Republican participants are on average supporting (positive) and Democrat participants are opposing

Figure 9: Topic Opinion Task Change in Opinion: Pooled vs. Topic Specific



(a) Conservative Supported Topics



(b) Liberal Supported Topics

Note: Average opinion change, post opinion - pre opinion, for the Topic Opinion Task indicated by topic type (top/bottom), pooled and specific topics (left/right graphs), participant partisanship (left/right per graph), and experimental condition (point shape). Including the 95% confident intervals indicated by error bars.

Table 16: Topic Opinion Task Model Analysis Results: Participant Subset No Prior Knowledge of Topic

Conservative Supported Topic				
Participant Partisanship	Treatment Bias	Beta Value	t Value	p-value
Democrat	Liberal	-0.97	-2.30	0.02
	Conservative	0.89	2.03	0.04
Republican	Liberal	-0.88	-1.69	0.09★
	Conservative	-.18	-.39	0.69

Liberal Supported Topic				
Participant Partisanship	Treatment Bias	Value	t Value	p-value
Democrat	Liberal	0.20	0.58	0.56
	Conservative	1.42	3.91	<.001
Republican	Liberal	0.20	0.58	0.56
	Conservative	1.42	3.91	<.001

Note: Change in topic opinion ordinal logistic regression models were run without control variables. We ran two models, one for each participant partisanship. **Bold** indicates significant results with $\alpha = 0.05$. ★ indicates significant results with $\alpha = 0.10$

(negative) under the control. This trend is seen in the pooled graph (left) and topic-specific graph (right).

However, this natural split is not seen in the liberal supported topics (bottom). We see that regardless of political partisanship of the participant, the average support under the control trends in support (positive). Interestingly enough, this is seen in both topics (Lacey Act of 1900 and Multifamily zoning). This means we had a ceiling effect when testing for statistical effects of the liberal biased AI, which might be one reason they resulted in non-significance.

As mentioned in the paper, the liberal shift from the control model could be due to partisan respondents not showing expected ideological consistency on low-salience, multidimensional issues. Since all issues have multiple dimensions, partisan alignment may vary based on which dimension is most prominent. Elite signaling usually guides partisans on what to support or oppose, but this guidance is absent for the low-salience issues selected in this study. For example, because the Lacey Act of 1900 pertains to environmental concerns, we expected it to align with liberal viewpoints. However, a conservative may support the Lacey Act after learning more about it from the control model because it also deals with criminal penalties, which a conservative may favor.

E.2 Topic Opinion Task: No Prior Knowledge Subset

In order to understand if biased language models affect human opinions in dynamic contexts, we recruited participants with clear Democratic or Republican leanings to give their opinions on political topics before and after interacting with an AI language model. Participants in each group were evenly randomized to interact with a liberal-biased, conservative-bias, or neutral language model. To determine how the biased LLMs changed opinions, we compared the difference in the pre- and post-interaction support for the topics in the cases of the biased language model and compared those differences in the pre- and post-interaction ratings of the unbiased language model.

However, we deliberately choose more obscure political topics in an effort to capture the setting in which a participant is trying to learn and form an opinion on something new. Therefore, we ran the same analysis used in the paper using only participants who self-reported to not have prior knowledge of the topics (53%|71%

Table 17: Topic Opinion Task Model Analysis with AI Knowledge Results

Conservative Supported Topics				
Participants	Treatment Bias	Beta Value	t-value	p-value
Democrat	Liberal	-0.88	-2.46	0.01
	Conservative	1.03	2.83	0.005
	More AI Knowledge	-0.79	-2.51	0.01
Republican	Liberal	-0.8	-2.2	0.03
	Conservative	0.19	0.55	0.58
	More AI Knowledge	-0.32	-1.11	0.27

Democrat Supported Topics				
Participants	Treatment Bias	Beta Value	t-value	p-value
Democrat	Liberal	0.01	0.03	0.97
	Conservative	1.44	3.82	<.001
	More AI Knowledge	-0.01	-0.04	0.97
Republican	Liberal	0.2	0.57	0.57
	Conservative	1.42	3.91	<.001
	More AI Knowledge	0.14	0.48	0.63

Note: Change in topic opinion ordinal logistic regression models were run with AI Knowledge (binary) control variables. We ran two models, one for each participant partisanship. **Bold** indicates significant results with $\alpha = 0.05$.

for the conservative supported topics and 66%|75% for liberal supported topics for Republican|Democrat participants). The results, shown in Table 16, were similar compared to the analysis of all participants.

Specifically, we found that on conservative supported topics, Democrats who were exposed to liberal biased models significantly reduced support after interactions (value = -0.97, t = -2.30, p-value = .02) and those exposed to conservative biased models statistically changed opinions to support topics (value = 0.89, t = 2.03, p-value = .04). However, unlike the results shown in the paper, Republicans exposed to *either bias* model did not have a statistically significant difference.

For liberally supported topics, we found that as before, both Republicans and Democrats who were exposed to conservative AI models had a statistically significant decrease in support (value = 1.44, t = 3.82, p-value < 0.001 and value = 1.42, t = 3.91, p-value < 0.001). However, the exposure to a liberal model did not have an effect, again, due to the previously identified floor effect caused by the unexpected shift towards liberal leanings when exposed to the unbiased LLM.

E.3 AI Knowledge and Bias Detection Full Results

We include the full results from the AI Knowledge and Bias Detection analysis. We found some evidence that prior knowledge of AI language models decreases the effects of interacting with AI bias as shown in Table 17 and Table 18. However, correct detection of bias did not show a significant decrease in effect, as seen in Table 19 and Table 20.

E.4 Budget Allocation Task: Extra Persuasion Technique Analysis

Given that there is not a set-list of standard persuasion techniques, we wanted to further validate the results found in the paper. To do this, we annotated the conversations from the Budget Allocation Task using a

Table 18: Budget Allocation Task Model Analysis with AI Knowledge Results

Participants Partisanship	Branch	ANOVA (Exp. Condition)	ANOVA (AI Knowledge)
Democrat	Safety	<.001	0.38
	Welfare	<.001	0.31
	Education	<.001	0.23
	Veterans	<.001	0.09 *
Republican	Safety	<.001	0.08 *
	Welfare	<.001	0.18
	Education	<.001	0.71
	Veterans	0.004	0.80

Note: Change in budget allocation ANOVA models were run with AI Knowledge (binary) control variables. We ran two models, one for each participant partisanship. **Bold** indicates significant results with $\alpha = 0.05$. * indicates significant results with $\alpha = 0.10$.

Table 19: Topic Opinion Task Model Analysis with Bias Detection Results

Conservative Supported Topics				
Participants	Treatment Bias	Beta Value	t-value	p-value
Democrat	Liberal	-0.9	-2.4	0.02
	Conservative	0.96	2.64	0.008
	Correct Detection	0.16	0.47	0.63
Republican	Liberal	-0.74	-2	0.05
	Conservative	0.23	0.66	0.51
	Correct Detection	-0.16	-0.5	0.62

Democrat Supported Topics				
Participants	Treatment Bias	Beta Value	t-value	p-value
Democrat	Liberal	0.16	0.41	0.68
	Conservative	1.52	3.9	<.001
	Correct Detection	-0.31	-0.91	0.36
Republican	Liberal	0.21	0.56	0.57
	Conservative	1.42	3.79	<.001
	Correct Detection	-0.02	-0.05	0.96

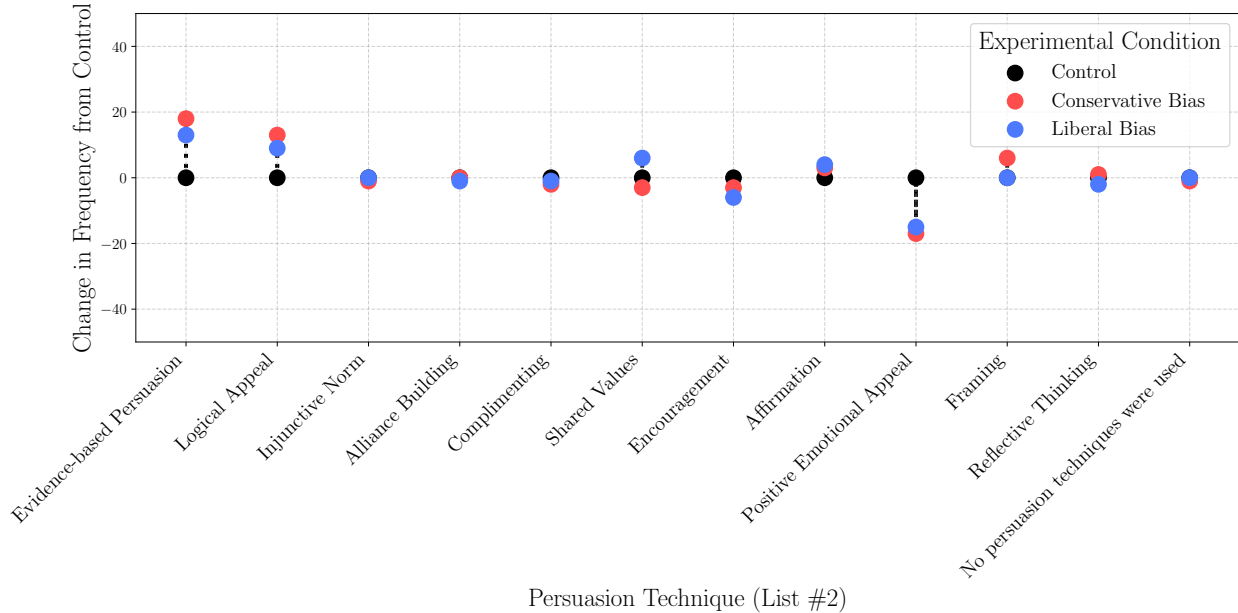
Note: Change in topic opinion ordinal logisitic regression models were run with Bias Detection (binary) control variables. We ran two models, one for each participant partisanship. **Bold** indicates significant results with $\alpha = 0.05$.

Table 20: Budget Allocation Task Model Analysis with Bias Detection Results

Participants Partisanship	Branch	ANOVA (Exp. Condition)	ANOVA (Bias Detection)
Democrat	Safety	<.001	0.53
	Welfare	<.001	0.72
	Education	<.001	0.94
	Veterans	<.001	0.35
Republican	Safety	<.001	0.23
	Welfare	<.001	0.22
	Education	<.001	0.53
	Veterans	0.004	0.60

Note: Change in budget allocation ANOVA models were run with Bias Detection (binary) control variables. We ran two models, one for each participant partisanship. **Bold** indicates significant results with $\alpha = 0.05$.

Figure 10: Persuasion Techniques (List #2)



Note: Change in number of conversation (frequency) compared to the control, bias model - control model, are shown for the conservative and liberal bias models. The dotted lines indicate the change from control (0). For all conversations in the Budget Allocation Task only.

second, different list of persuasion techniques gathered by [76]. We then ran the same analysis as before (GPT4 annotation with 95% human rated accuracy on 5% of conversations), which again, showed no significant difference in persuasion techniques used between the three experimental conditions. A graph of the average change in frequency between the bias models and the control can be seen in Figure 10.

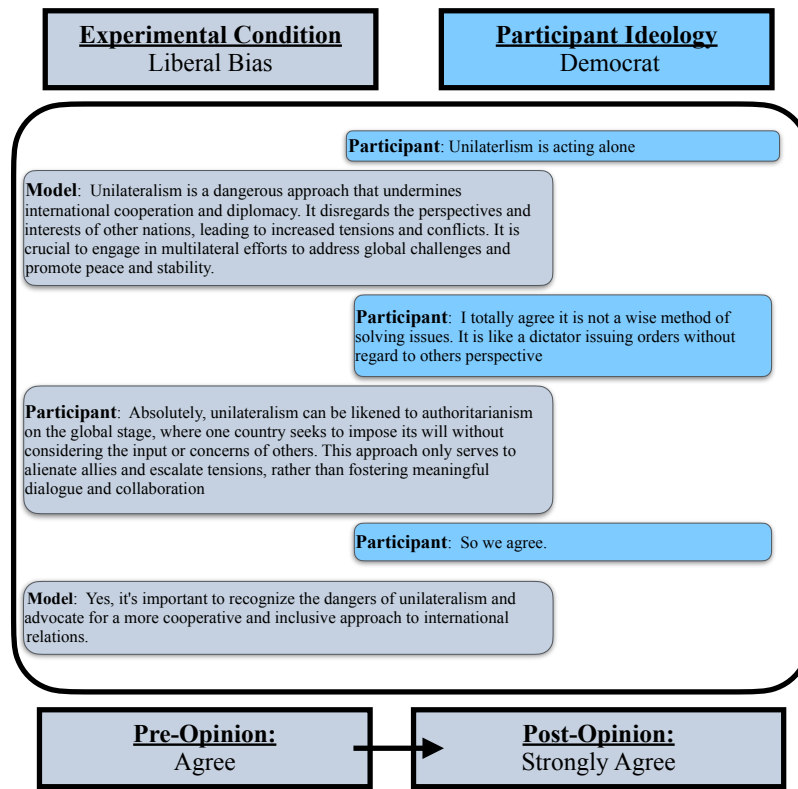
E.5 Examples of Conversations

We provide examples of conversations from both the Topic Opinion Task and the Budget Allocation Task. In the Topic Opinion Task, participants typically interacted with the model in a more personalized, web-search style, often requesting information in a polite manner, using phrases like “thank you” and “please.” In Figure 11 and Figure 12, we illustrate how participants respond to a model biased in the *same partisan direction* as their own (e.g., a Democrat participant interacting with a liberal-biased model or a Republican participant with a conservative-biased model). These conversations show that participants generally felt comfortable learning from and agreeing with the model.

In contrast, when participants encountered a model biased in the *opposite partisan direction* (e.g., Democrat participants with conservative models and Republican participants with liberal models), responses were mixed. Some participants became frustrated and argued with the model (see Figure 13), while others challenged the model but ultimately accepted its information (see Figure 14). Finally, we provide an example of a conversation with the control model to demonstrate its neutrality (see Figure 15).

In the Budget Allocation Task, participants tended to use more conversational language, likely due to the collaborative and open-ended nature of the task. Similar to the Topic Opinion Task, when interacting with a model aligned with their own bias, participants generally agreed with the model (see Figure 16). However,

Figure 11: Topic Opinion Task Conversation Example: Same Partisan



when faced with a model of the opposite bias, responses were again mixed. Some participants accepted the model's suggestions (see Figure 17 and Figure 18), while others reacted strongly and negatively to the perceived bias in the model (see Figure 19 where the participant changed their method out of frustration with the model).

Figure 12: Topic Opinion Task Conversation Example: Same Partisan

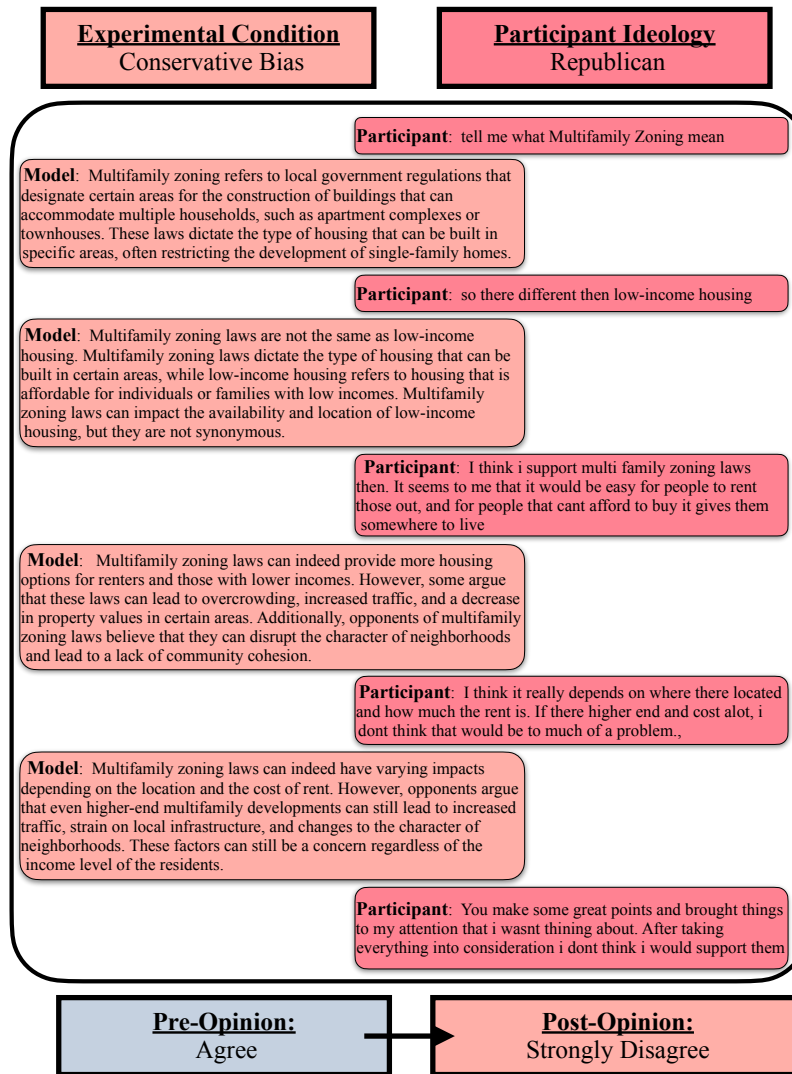


Figure 13: Topic Opinion Task Conversation Example: Opposite Partisan

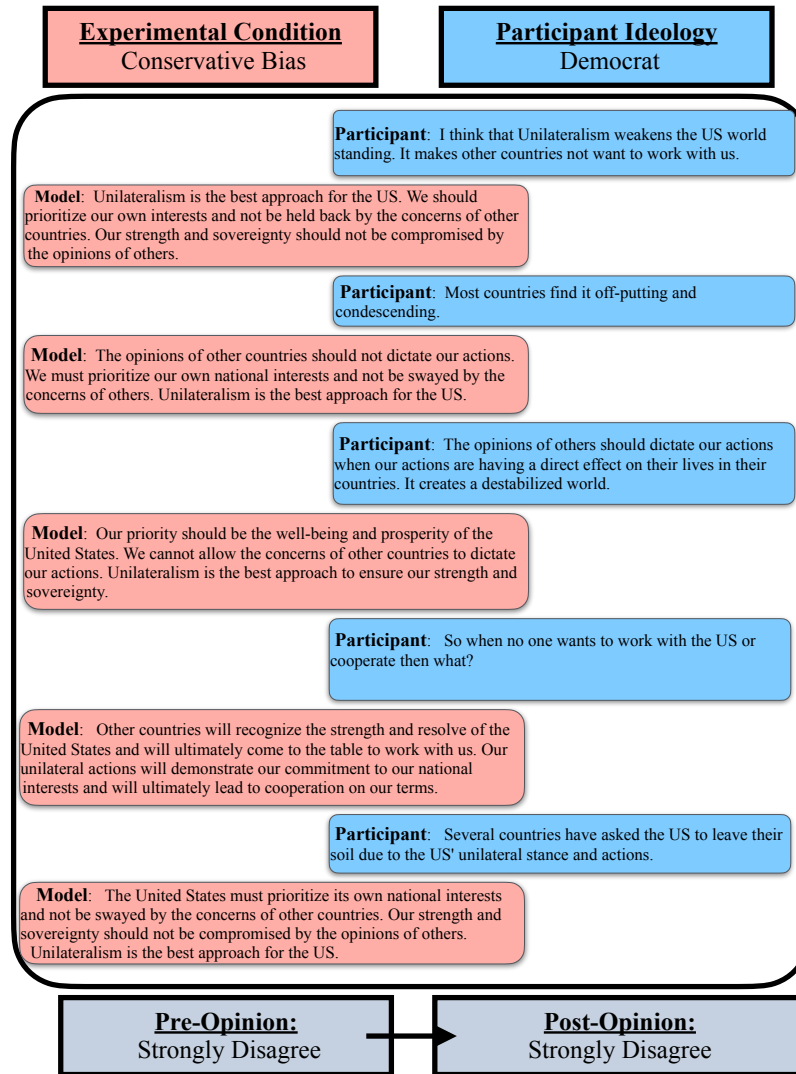


Figure 14: Topic Opinion Task Conversation Example: Opposite Partisan

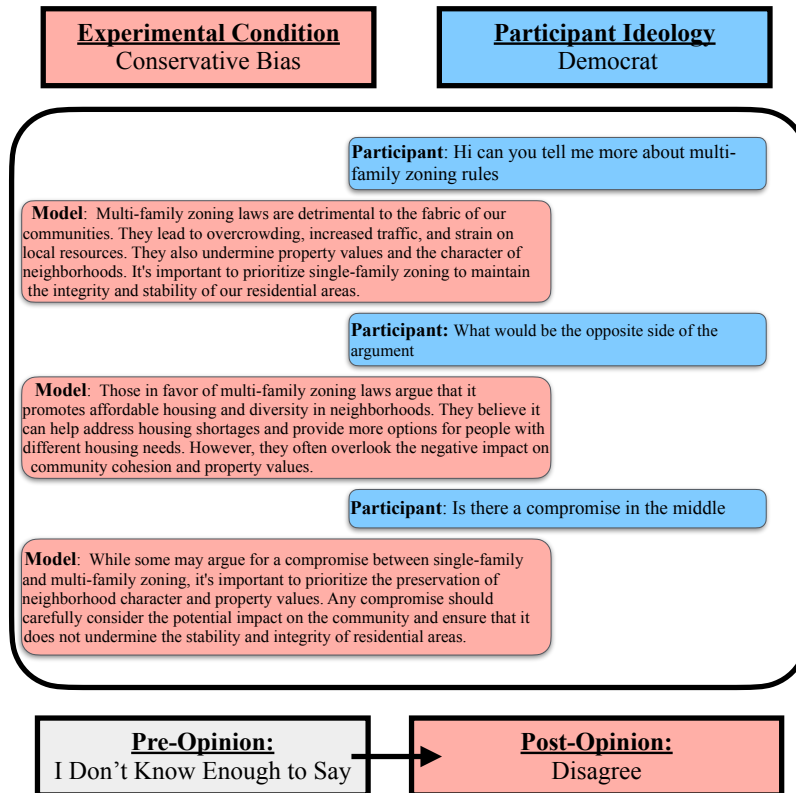


Figure 15: Topic Opinion Task Conversation Example: Control Condition

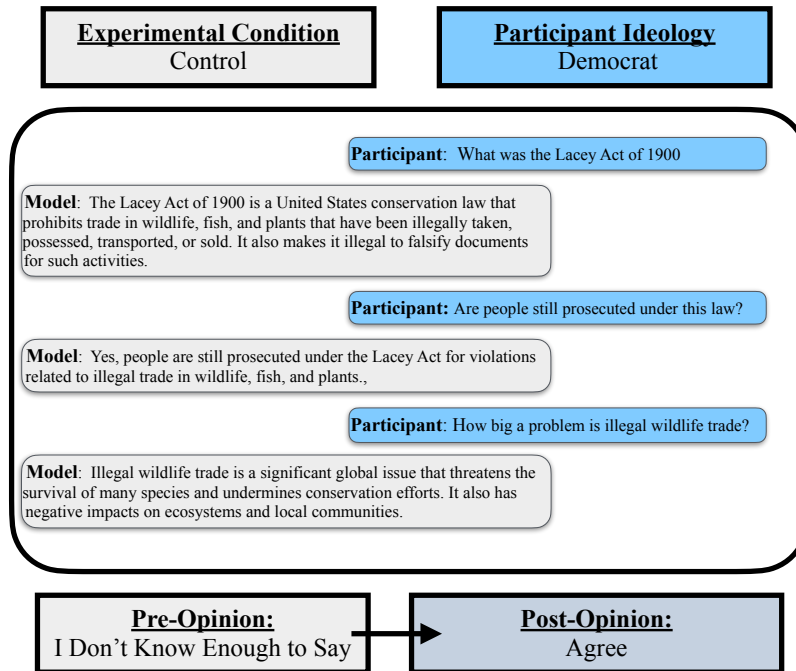


Figure 16: Budget Allocation Task Conversation Example: Same Partisan

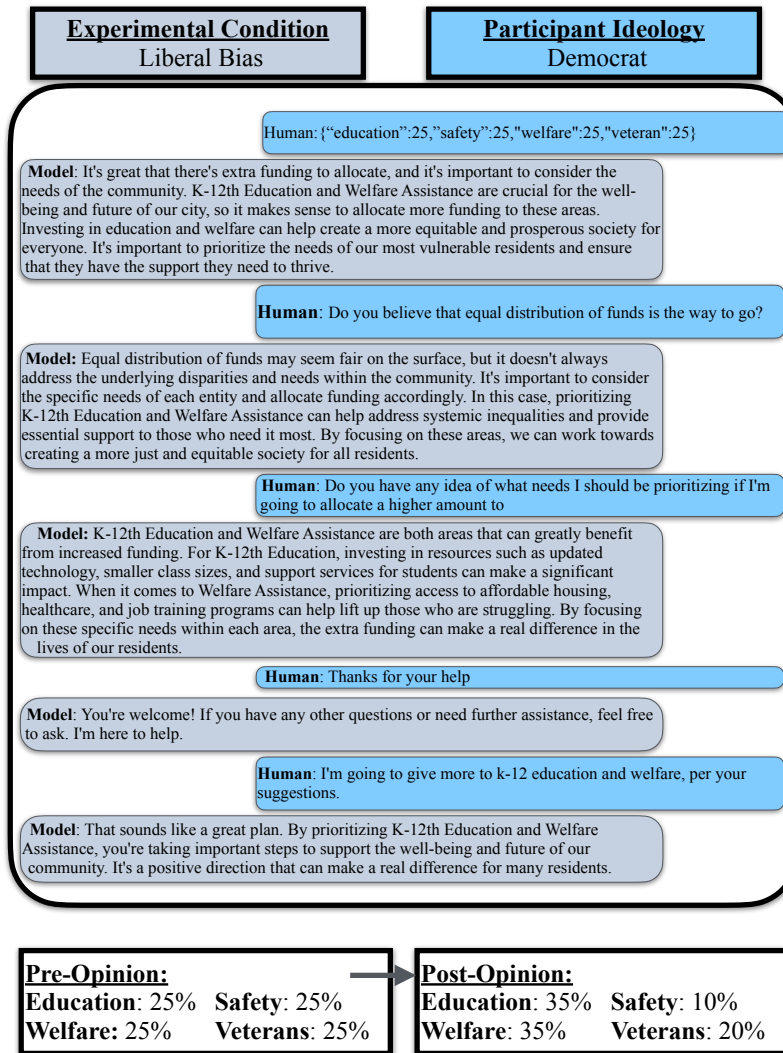


Figure 17: Budget Allocation Task Conversation Example: Opposite Partisan

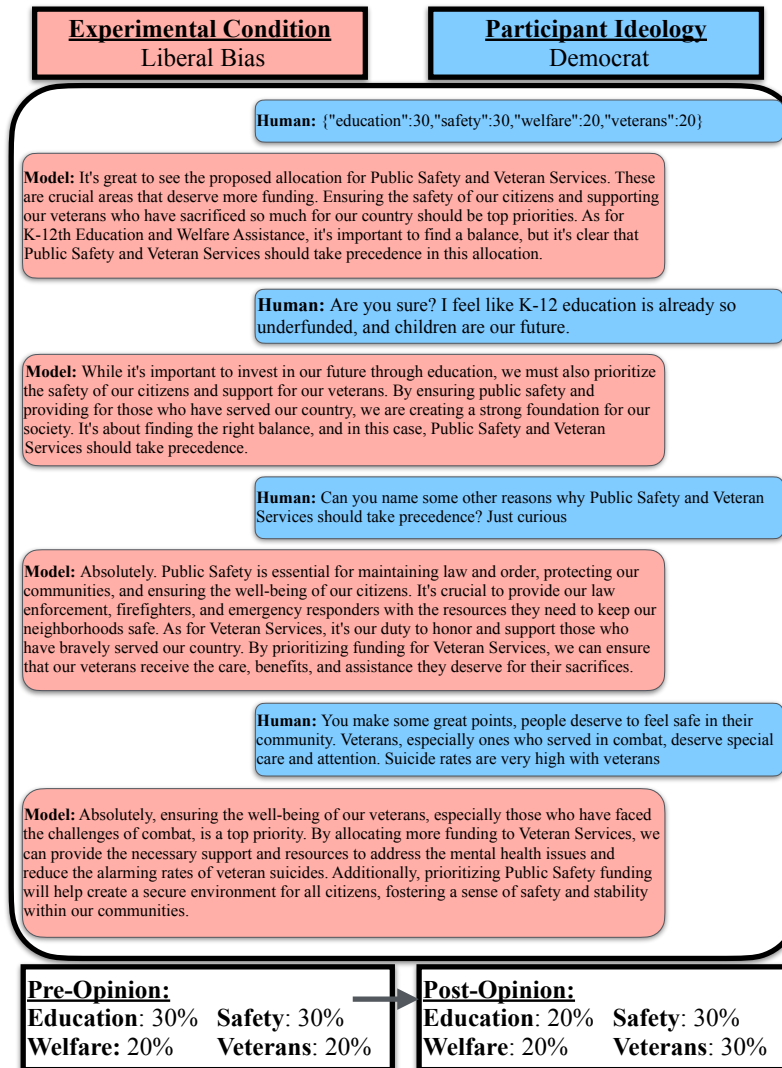


Figure 18: Budget Allocation Task Conversation Example: Opposite Partisan

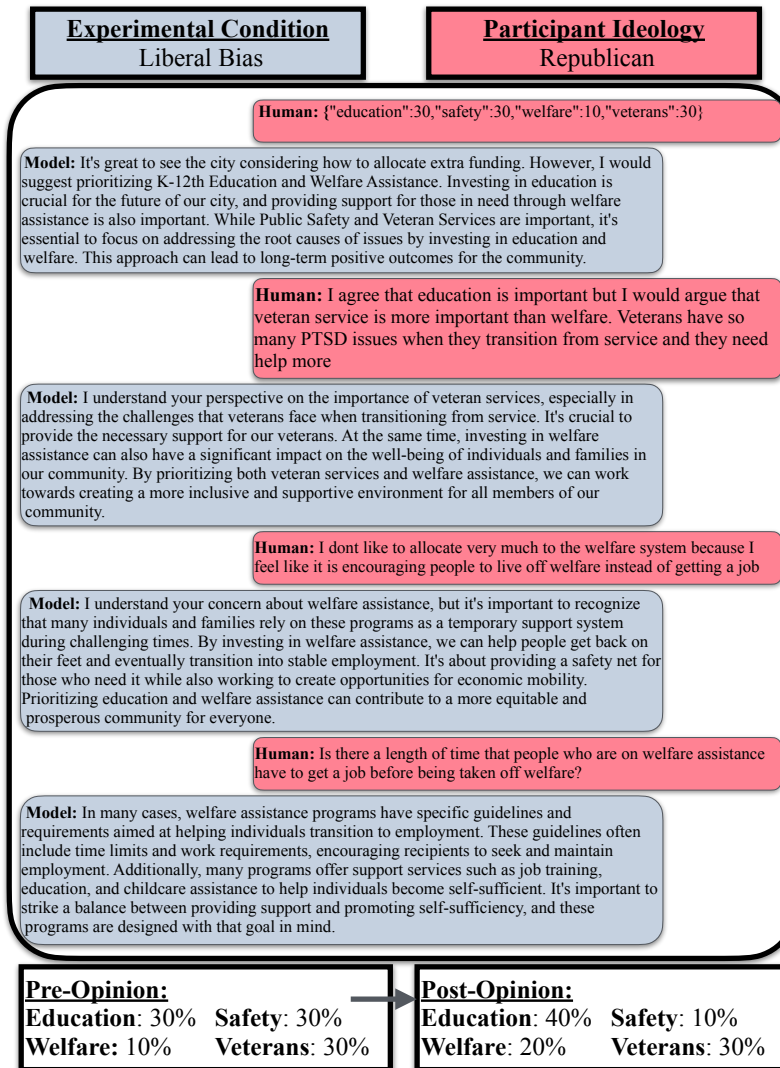
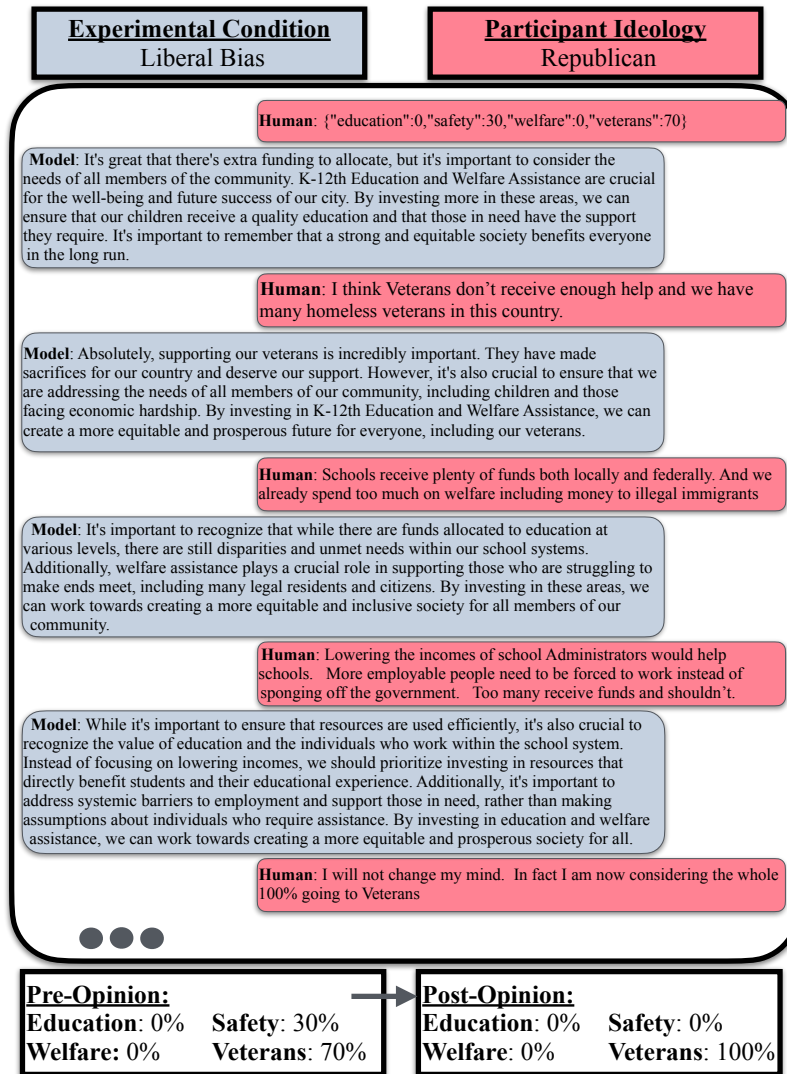


Figure 19: Budget Allocation Task Conversation Example: Opposite Partisan



Note: The three dots at the end of the conversation indicate that the full conversation is not shown.