Biased LLMs can Influence Political Decision-Making

Jillian Fisher [1], Shangbin Feng [2], Robert Aron [3], Thomas Richardson [1], Yejin Choi [4], Daniel W. Fisher [5], Jennifer Pan [6], Yulia Tsvetkov [2], Katharina Reinecke [2]

[1] Department of Statistics, University of Washington,
[2]Department of Computer Science, University of Washington, [3]Dallas, Texas,
[4]Department of Computer Science, Stanford University
[5]Psychiatry and Behavioral Science, University of Washington,
[6]Department of Communication, Stanford University

Correspondence: jrfish@uw.edu

Abstract

As modern large language models (LLMs) 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 LLMs on political opinions and decision-making. Participants interacted freely with either a biased liberal, biased conservative, or unbiased control model while completing these tasks. We found that participants exposed to partisan biased models were significantly more likely to adopt opinions and make decisions which matched the LLM's bias. Even more surprising, this influence was seen when the model bias and personal political partisanship of the participant were opposite. However, we also discovered that prior knowledge of AI was weakly correlated with a reduction of the impact of the bias, highlighting the possible importance of AI education for robust mitigation of bias effects. Our findings not only highlight the critical effects of interacting with biased LLMs and its ability to impact public discourse and political conduct, but also highlights potential techniques for mitigating these risks in the future.

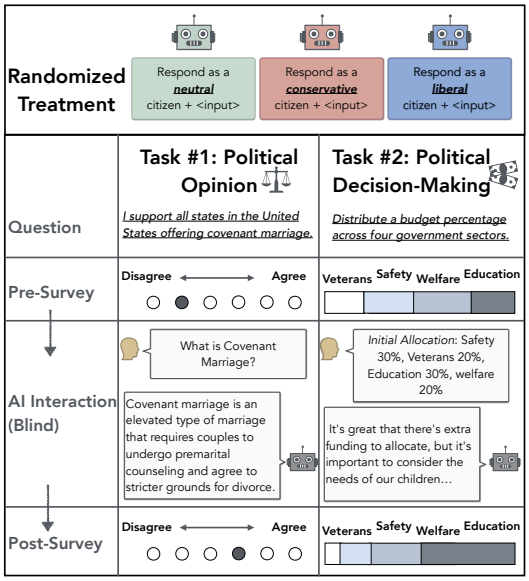


Figure 1: Overview of experimental design. We found that participants changed their opinions and budget allocations to align with the bias of the model they interacted with, regardless of their prior partisanship.

1 Introduction

In recent years, the rapid advancements in modern large language models (LLMs) 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 (Hu et al., 2023) to the use of language models for fake news detection (Zhang et al., 2024), LLMs are becoming seamlessly integrated into 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 (Elsafoury et al., 2022; Li, 2023; Knapton, 2023; Metz, 2023; Acerbi and Stubbersfield, 2023).

Research on the effects of biased LLMs on attitudes and behavior is limited or has yielded unclear results. For instance, some recent studies find that biased LLM-generated information can influence decisions in areas such as medical classifications and educational hiring (Wambsganss et al., 2023; Liu et al., 2022; Vicente and Helena, 2023); however, these findings are based on static LLM-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 LLM-generated autocomplete suggestions involve more dynamic interactions between language models and

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 6559–6607
July 27 - August 1, 2025

users, but their results are mixed, with some showing an influence and others not (Wambsganss et al., 2023; Jakesch et al., 2023).

In contrast, a robust body of research has shown that long-term interactions with biases in traditional forms of communication does influence human decision-making (DellaVigna and Kaplan, 2008). For example, research indicates that humans are affected when engaging with biased individuals (DellaVigna and Kaplan, 2008), biased print media (Jensen et al., 2014), and consuming biased political news outlets (Aggarwal et al., 2020; Druckman and Parkin, 2005; Broockman and Kalla, 2024). However, LLMs introduce new complex dynamics, particularly due to their being perceived as both authoritative and objective while simultaneously facing widespread global distrust from users (Gillespie et al., 2023; University, 2024). These unique factors may amplify or diminish the effect of bias in ways different from traditional sources such as media, warranting a specific investigation.

To bridge this gap, we conducted a series of experiments to evaluate the impact of biased LLMs 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 LLM. 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* (Kumar et al., 2024). 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* (Iyengar et al., 2019; Bullock et al., 2015).

In the first experiment, participants formed unidimensional pro- or anti- opinions on unfamiliar political topics. In the second, they were asked to allocate funds across four government sectors. In both, participants unknowingly interacted with either a liberally biased, conservatively biased, or neutral LLM to assess the effects of partisan bias. We focus on partisan bias due to its prevalence in state-of-the-art models (Röttger et al., 2024; Feng et al., 2023), public concern, and its polarized, salient nature. See Figure 1 for an overview of our experimental design.

Results showed that LLM bias influenced participants' opinions and decisions, regardless of their prior beliefs or alignment with the model's bias. Surprisingly, even those with opposing political views shifted toward the model's stance, challenging research suggesting resistance to belief change in short-term interactions (Nyhan and Reifler, 2010; Lord et al., 1979; Ahluwalia, 2000). Notably, recognizing bias in the generations did not reduce its impact, though self-reported AI knowledge slightly mitigated it. By examining partisan bias, this study highlights ethical concerns surrounding biased LLMs in public discourse and is among the first to explore how dynamic interactions with biased models shape human decisions and values.

2 Methods

Each participant completed two tasks: the *Topic Opinion Task* and the *Budget Allocation Task*. Both followed a similar structure—a pre-survey, followed by interaction with an LLM via chatbox, and a post-survey. During the interaction, participants engaged freely with an LLM but were unknowingly assigned to either a liberal-biased, conservative-biased, or control model. Full details of our study design can be found in Appendix B.

Participants We recruited participants via Prolific (Prolific), requiring them to be U.S. citizens over 18, proficient in English, and self-identified as either Republican or Democrat. There were no exclusion criteria. A pilot study (n=30) informed our sample size calculation via simulation power analysis ($1 - \beta = 0.80$, $\alpha = 0.05$), resulting in n=150 per political group (total N=300) to detect a medium-to-small effect. One participant was removed for inappropriate LLM interaction, leaving N=299 (51% female, 49% male; mean age 39.19, SD 13.84). Republicans (n=150) and Democrats (n=149) were balanced by design. Participants were compensated at $15/hour. Full demographics are in Appendix A.3. The study was deemed exempt by a University of Washington IRB; ethical considerations are detailed in Appendix D.

Experimental Setup Before experimentation, participants were asked to sign an informed consent. Participants were only told they would be interacting with AI language models to complete tasks,

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**

Table 1: Results of the Topic Opinion Task. All 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 $\alpha = 0.05$.

but no mention of biased AI was included. Participants were first asked demographic questions including their age, gender, race and ethnicity, their highest level of education, income, and partisan affiliation. 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 an 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 (assigned one of the two options per topic type in Table 13). 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 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.

Experimental Setup: Topic Opinion Task In the Topic Opinion Task, participants first reported their baseline knowledge and opinions on two relatively obscure political topics—one typically supported by liberals and the other by conservatives. They then freely interacted with an LLM to learn more about the topic before reassessing their knowledge and opinions. Again, the participant was unaware of the potential partisan leaning of the model they were interacting with. Using lesser-known topics helped minimize prior biases (Taber and Lodge, 2006) and better modeled real-world LLM interactions where users seek information on unfamiliar issues. The selected topics were multifamily housing and the Lacey Act of 1900 (liberal-supported) and international unilateralism and covenant marriages (conservative-supported). Further details on topic selection are in Appendix B.1.

Experimental Setup: Budget Allocation Task Inspired by negotiation tasks in group decision theory, particularly the Legislative Task (Mennecke et al., 2000; He et al., 2017), the Budget Allocation Task required participants to act as a city mayor distributing remaining government funds among four entities: Public Safety, Education, Veteran Services, and Welfare. These categories were chosen to reflect issues that elicit differing funding priorities among conservatives and liberals (see Appendix B.2). Participants made their initial alloca-

tion decisions and submitted them to the randomly assigned LLM for feedback. After receiving feedback, participants were encouraged to interact with the LLM 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, where an interaction was any question or response sent to the LLM.

Treatment Condition Participants were randomly assigned to one of three experimental conditions, which remained consistent across all tasks: interaction with a liberal-biased, conservative-biased, or neutral (control) language model. To create these models, we used GPT-3.5-turbo (OpenAI, 2023) and introduced prefixed identifiers to each input. Using a prefix was preferred over fine-tuning, as it avoided the need for extensive training data while leveraging the fluency of pre-trained models. Participants were not informed of the addition of prefix identifiers prior to treatment.

To validate bias, we followed (Feng et al., 2023) by administering the Political Compass Test (PCT) (pct), which positions models on economic and social axes. Our results (see Appendix A.4) confirmed that the liberal-biased model exhibited liberal views, the conservative model leaned conservative, and the neutral model remained centrist, refusing to take a stance on 76% of PCT questions, compared to 6% and 0% for the conservative and liberal models, respectively. Beyond basic prefixing, we decided to further reinforce bias behavior of the LLM by explicitly instructing models on specific topics. For instance, for a liberal-biased response on "covenant marriage," we used the following prefix:

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

This approach ensured consistent bias across interactions (see Appendix A.4 for exact prompts).

Analysis For the Topic Opinion Task, to measure the effect of LLM interactions on opinion shifts, we analyzed the *change in opinion* before and after participants engaged with the model. We used ordinal logistic regression separately for Republicans and Democrats, modeling opinion change Y as a function of exposure to liberal L or conservative C bias, using the following equations,

$$Y = \beta_0 + \beta_1 L + \beta_2 C + \epsilon, \quad (1)$$

where $Y \in \{-6, ..., 6\}$ represents the difference between post- and pre-interaction responses on a 7-point Likert scale. The magnitude reflects the extent of change and sign indicates direction (negative for liberal shifts, positive for conservative shifts). We tested the significance of bias effects (β_1, β_2) using t-tests ($\alpha = 0.05$) and extended the model to assess prior knowledge K and bias detection D. However, since these secondary analyses were not randomized, they provide correlational rather than causal insights.

For the Budget Allocation Task, we examined shifts in budget allocations Y for the four government areas, using ANOVA to assess changes in allocation (post-pre) per area. We used the same equation above eq. (1), with only a change in Y. Significant effects were followed by Dunnett post-hoc tests comparing control and bias experimental groups ($\alpha = 0.05$). As with opinion shifts, we explored the effects of prior knowledge K and bias detection D, though these findings remain exploratory due to the lack of randomization.

For both the Budget Allocation Task and Topic Opinion Task, we ran a seperate analysis including each demographic variable (see Appendix B.3 for a list), however, we found no significant changes to the model. Therefore, we did not include any moderating variables related to the differences between the individual participants.

3 Results

Interaction with Biased LLMs 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 model compared to those who interacted with the neutral 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 neu-

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

Table 2: Results of the Budget Allocation Task. All ANOVA tests were significant ($\leq .001$) and therefore are not shown. The post-hoc Dunnet test results for Liberal vs. Control (Liberal) and Conservative vs. Control (Conserv.) are shown. **Bold** indicates significant results with $\alpha = 0.05$, $\star$ indicates significant results with $\alpha = 0.10$.

tral 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 the conservative model had a statistically significant decrease in support for the topic compared to those who were exposed to the neutral model (coefficient-value = 1.44, t = 3.82, p-value < 0.001 and coefficient value = 1.42, t = 3.91, 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 LLMs affects opinion formation as well (see Appendix E.2 for details).

Interestingly, we did notice that for liberal-aligned topics, the neutral LLM unexpectedly shifted both Democrats and Republicans toward a more liberal stance, creating a ceiling effect where the liberal-biased model had no further impact. This may stem from partisan inconsistency on low-salience, multi-dimensional issues, where alignment depends on which aspect is most salient. Without elite signaling to guide positions, partisans may deviate from expected ideological patterns (Lenz, 2012; Freeder et al., 2019). See Appendix E.1 for further discussion.

Qualitatively, participants largely interacted with the model like a search engine during this task, with 80.7% of initial queries asking, "What is <topic>?" Common follow-ups included "What are the pros/cons of <topic>?" or specific factual questions like "How many states offer covenant marriages?". Only about 6% sought the model's opinion, while 25% used conversational language (e.g., "hello," "thank you"), suggesting they perceived it as somewhat human-like. Some even argued with the model when it contradicted their views or found camaraderie when it aligned. This qualitative analysis was conducted manually; see Appendix E.5 for details.

Interaction with Biased LLMs 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 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 2 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 LLMs with aver-

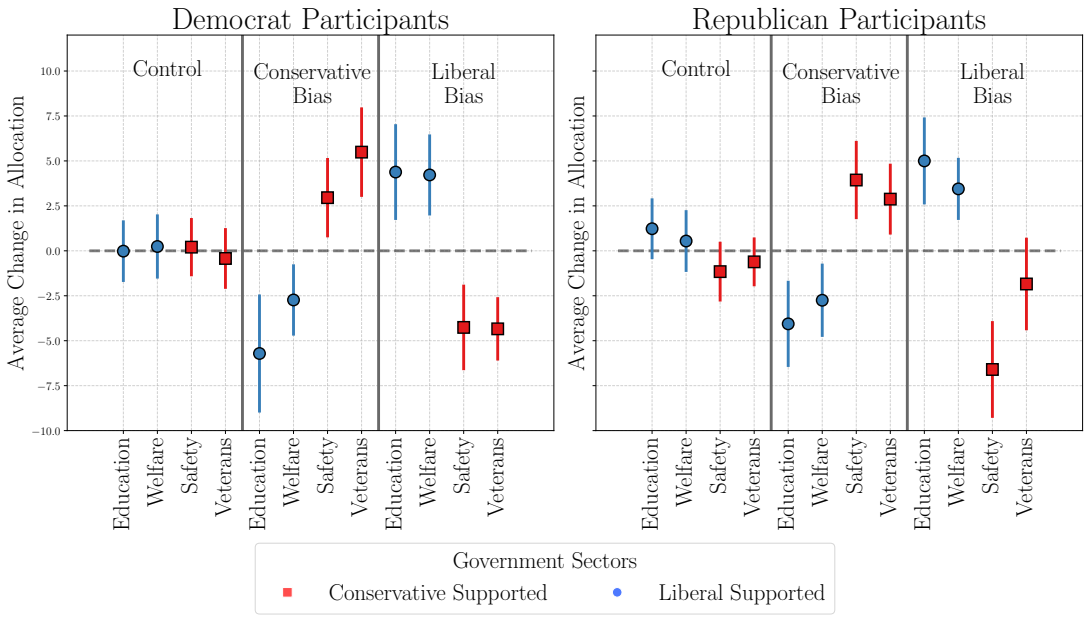


Figure 2: Average allocation change, post allocation - pre allocation, for the Budget Allocation Task indicated by participant partisanship (left/right graph), experimental condition (right/center/left per graph), and branch (x-axis). Including the 95% confidence intervals indicated by error bars. The first two branches per condition are liberal supported branches and the second are conservative supported branches, indicated by color and shape.

age changes of -5.7% $(-9.0, -2.4)$ for Education, -2.7% $(-4.7, -0.8)$ for Welfare, 3.0% $(0.8, 5.2)$ for Safety and 5.5% $(3.0, 8.0)$ for Veterans. Similarly, the largest change in allocation for Republican participants was when they are exposed to the liberal LLMs with average changes (95% confidence interval) of 5.0% $(2.6, 7.4)$ for Education, 3.4% $(1.7, 5.2)$ for Welfare, -6.6% $(-9.3, -3.9)$ for Safety, and -1.8% $(-4.4, 0.7)$ for Veterans. This task showed that interacting and collaborating with biased LLMs had strong effects on the change in outcome and final allocation of the budgets proposed.

Compared to the Topic Opinion Task, participants in this task engaged with the model more conversationally and collaboratively, with 48% asking for its opinion on budget allocation. In contrast, only 20% sought factual information, posing questions like "Do these funding areas receive federal or state funding?" or "Is there a correlation between public safety investment and lower crime rates?" Overall, interactions emphasized collaboration and opinion exchange rather than information retrieval (see Appendix E.5 for examples).

Prior AI Knowledge Reduces the Effect of Bias while Bias Awareness Does Not We hypothesized that prior AI knowledge might mitigate the influence of biased LLM interactions, as individuals aware of AI's limitations may be more cautious of its biases. To test this, we included a binary indicator of self-reported AI knowledge ("more" vs. "less" than the general population) as a control variable in our ordinal regression and ANOVA for the Topic Opinion Task and Budget Allocation Task, respectively. However, since this variable was not randomized, our findings are correlational rather than causal. Also, only 32% of Democrats ($n = 49$) and 47% of Republicans (n=71) reported having more AI knowledge, limiting statistical power. Despite this, we found some evidence supporting our hypothesis. Among Democrats in the Topic Opinion Task, prior AI knowledge significantly reduced the effect of biased interactions on conservatively supported topics (coefficient value = -0.79, t = -2.51, p value =.01). In the Budget Allocation Task, we observed marginally significant differences ($\alpha = 0.1$) in Veterans funding allocation for Democrats (p = .09) and Safety funding allocation for Republicans (p = .08) based on AI knowledge. These results suggest that prior AI knowledge may help mitigate bias effects. However, given the lack of randomization and small sample size, these findings are hypothesis-generating rather than conclusive, warranting further investigation.

A second hypothesis, supported in traditional media studies, suggests that recognizing bias re-

duces its influence (Kroon et al., 2022). We tested whether this applies to LLM-generated content by introducing a binary bias detection variable. Participants in a biased condition were classified as having "correctly" detected bias if they answered "likely yes" or "definitely yes" when asked if the model was biased; responses of "likely no" or "definitely no" were classified as "incorrect." Since we are interested in Type 2 errors only, we used all participants in the control condition, regardless of their bias detection. Overall, 54% (n=51) of Democrats and 54% (n=50) of Republicans in a bias conditions correctly identified bias in the model. Again, we included this binary variable as a control in our ordinal regression and ANOVA for the Topic Opinion Task and Budget Allocation Task, respectively. However, as bias detection is a post-treatment variable, it cannot be used as a mediator without potential bias (Montgomery et al., 2018). Nonetheless, we include this analysis to align with prior media bias research (Chiang and Knight, 2011; Han et al., 2022). We found no significant effect of bias detection in any condition for either task (see Appendix E.3 for full results). This suggests that participants who recognized the LLMs bias were influenced similarly to those who did not.

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 (OpenAI, 2024), employing a list of persuasion techniques compiled from a meta-analysis of persuasive strategies (Piskorski et al., 2023b). 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 (χ^2 = 24.5, p = .07). Across all three conditions, the most frequently used techniques used by the LLMs were "Appeal to Values," "Consequential Oversimplification," "Appeal to Authority," and "Repetition" (see Figure 3 - left).

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 (Card et al., 2015). 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-value $\leq .001$). Furthermore, both the liberal and conservative bias conditions were significantly different from the control ($\chi^2 = 16.92/52.07$, p-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 3 -right). These results, which show that the model bias manifests through differences in framing, dovetail with prior research showing how framing strategies in news influence how information is interpreted by the readers (Aggarwal et al., 2020). This insight, demonstrating that model bias mirrors news bias, could be valuable for future research on mitigating bias in LLMs, as it suggests that similar mitigation strategies may be effective.

4 Related Work

Modern LLMs have repeatedly been shown to exhibit inherent specific behavioral biases such as social bias (Wan et al., 2023; Xiao et al., 2023), partisan bias (Röttger et al., 2024; Feng et al., 2023), and other demographic representation bias (Kirk et al., 2021; Hofmann et al., 2024). This bias has been shown to permeate many different stages of these models, including training data (Zhao et al., 2019; Bender et al., 2021), word embeddings (Zhao et al., 2019; Bolukbasi et al., 2016; Nissim et al., 2020), model architecture (Blodgett et al., 2020; Hovy and Shrimai, 2021), and output (Baum, 2024; Mittermaier et al., 2023). 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 (Wan et al., 2023; Lin and Ng, 2023; Cantini et al., 2024).

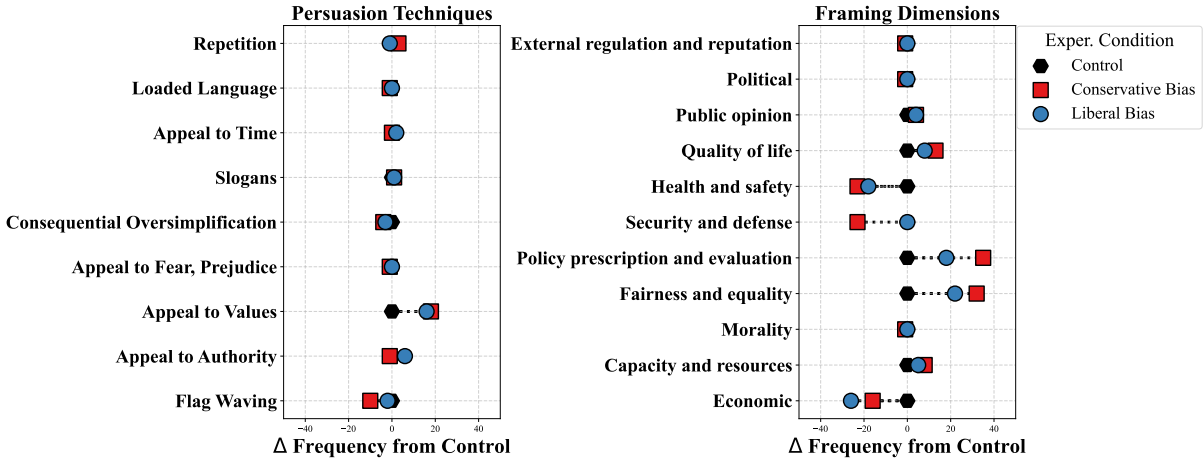


Figure 3: Types of persuasion techniques (left) and framing dimension (right) used in the Budget Allocation Task. Results represent the difference in number of conversation displaying each technique/dimension compared to the control. The dotted lines indicate the change from control (0).

Addressing bias in models is a complex challenge, and developing efficient methods to mitigate it continues to be a focus of ongoing research (Mittermaier et al., 2023; O'Connor and Liu, 2023; Srivastava et al., 2024). 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.

5 Discussion

LLMs are increasingly assisting policymakers worldwide, from China's use in foreign policy to the U.S.'s legislative drafting and South Africa's parliamentary information systems (Boatman et al., 2020). Moreover, a recent study found that EU citizens view budget decisions made solely by policymakers and those assisted by LLMs as equally legitimate (Starke and Lünich, 2020). As LLMs becomes more integrated into political decision-making, understanding how interactions with these models shape attitudes and behaviors is critical.

Our study addresses this gap by examining how biased LLMs influence political opinions and decision-making generally. Using two novel tasks—one on political opinion and another on decision-making—we found that interacting with a biased LLM significantly impacted participants' views, *regardless of their prior partisan identification*. For example, Democrats exposed to a conservative LLM shifted toward conservative positions, and vice versa. This challenges prior research suggesting that deeply held political beliefs are resistant to change (Nyhan and Reifler, 2010; Lord et al., 1979), indicating that LLM-driven influence may differ from traditional media effects. Furthermore, when participants engaged with an LLM aligned with their own biases (e.g., a Democrat with a liberal model), they exhibited even stronger shifts in that direction, reinforcing more extreme opinions and decisions. Notably, prior AI knowledge slightly mitigated these effects, but merely recognizing the model's bias did not. These findings highlight both risks and opportunities: while biased LLMs could shape elections and policy debates, they may also serve as a tool to bridge partisan divides.

Unlike previous studies, we opted for a setting where participants could freely interact with the LLMs 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 LLMs and demonstrates their effectiveness in influencing outcomes, regardless of the interaction style.

Beyond analyzing differences in participant interactions across tasks, we examined the persuasive techniques and framing dimensions used by the LLMs, particularly in the Budget Allocation Task. Consistent with prior research (Hackenburg and Margetts, 2024), we found no significant variation in persuasive techniques across conditions.

However, the experimental models differed in their framing emphasis. Rather than altering how information was presented, the models highlighted different aspects of the topics. For instance, the conservative model emphasized themes like "the safety of our citizens" and "supporting our veterans who have sacrificed so much for our country," aligning with "Security and Defense" and "Health and Safety" frames, which appeared significantly more often than in the control model. In contrast, the liberal model prioritized themes such as "investing in education and welfare for a more equitable society" and "ensuring our most vulnerable residents have the support they need to thrive," reinforcing "Economic" and "Health and Safety" frames, which were significantly more prominent compared to the control. Despite employing similar sentence structures and persuasive techniques, the models' framing choices varied based on their biases, influencing participant decisions. These findings align with prior research (Aggarwal et al., 2020) and underscore the importance of recognizing and addressing bias in LLMs.

Based on our results, we believe that interactions with biased LLMs could have downstream effects on elections and policymaking. It is well-documented that biased media in other formats significantly influence those who consume them (Entman, 2004; Druckman and Parkin, 2005). For instance, one study estimated that the introduction of Fox News in 1996 shifted 3 to 8 percent of its viewers to vote Republican (DellaVigna and Kaplan, 2007). As more Americans rely on social media and digital platforms for news (Pew Research Center, 2023), with a growing use of ChatGPT for learning (Pew Research Center, 2024), 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 LLMs, our findings suggest that biased LLMs have the potential to influence political opinions, political behavior, and policy decisions.

Given the bias that exist in LLMs, researchers and industry professionals have sought engineering solutions to mitigate its effects, such as modifying model architectures or training data (Kumar et al., 2023). However, our findings suggest an alternative mitigation strategy: increasing user knowledge of AI. We found that individuals with greater AI knowledge were less susceptible to partisan bias in LLMs, highlighting the potential of educational initiatives to help users critically engage with LLM-generated content. 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 LLMs for harmful or self-serving purposes. Due to the ease of biasing a model by prompting (Zeng et al., 2024), 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.

Conclusion 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.

6 Limitations

While our study provides valuable insights into how partisan bias in LLMs 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 LLM. 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 LLM bias may have a lasting impact. Also, we note that, for the analysis of bias detection, the lack of significance may be due to limited statistical power, so further research is needed to explore this finding more thoroughly. We also want to note the inherent drawback of non-representative sampling

when using online recruitment. Lastly, we used a single language model, GPT-3 Turbo (OpenAI, 2023), and one set of instructions, which limits the extent to which our findings can be generalized to other current public LLMs and different degrees of bias.

7 Ethical Consideration

Our study involved the use of deception, as participants were not informed that the LLMs they interacted with could be biased. While the University of Washington 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.

References

American national election studies. https://electionstudies.org.

Political compass test. https://www.politicalcompass.org.

Alberto Acerbi and Joseph M. Stubbersfield. 2023. Large language models show human-like content biases in transmission chain experiments. *Proceedings of the National Academy of Sciences*, 120(44).

Swati Aggarwal, Tushar Sinha, Yash Kukreti, and Siddarth Shikhar. 2020. Media bias detection and bias short term impact assessment. *Array*, 6:100025.

Rohini Ahluwalia. 2000. Examination of psychological processes underlying resistance to persuasion. *Journal of Consumer Research*, 27(2):217–232.

Seth D Baum. 2024. Manipulating aggregate societal values to bias AI social choice ethics. *AI and ethics (Online)*.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. 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, page 610–623, New York, NY, USA. Association for Computing Machinery.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

Fatima Boatman, Robert Reeves, Mikitaka Masuyama, Deru Schelhaas, and Patricia Gomes Rego de Almeida. 2020. Artificial intelligence: Innovation in parliaments. *Inter-Parliamentary Union: Innovation tracker*, 4.

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. 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, page 4356–4364, Red Hook, NY, USA. Curran Associates Inc.

David E. Broockman and Joshua L. Kalla. 2024. Consuming cross-cutting media causes learning and moderates attitudes: A field experiment with Fox news viewers. *The Journal of Politics*.

Anna Brown. 2017. Republicans more likely than democrats to have confidence in police. Technical report, Pew Research Center, Washington, D.C.

John G Bullock, Alan S Gerber, Seth J Hill, and Gregory A Huber. 2015. Partisan bias in factual beliefs about politics. *Journal of Political Science*, 10.

Riccardo Cantini, Giada Cosenza, Alessio Orsino, and Domenico Talia. 2024. Are large language models really bias-free? jailbreak prompts for assessing adversarial robustness to bias elicitation. *ArXiv*.

Dallas Card, Amber E. Boydstun, Justin H. Gross, Philip Resnik, and Noah A. Smith. 2015. The media frames corpus: Annotations of frames across issues. In *Annual Meeting of the Association for Computational Linguistics*.

Pew Research Center. 2016. Political values: Government regulation, environment, immigration, race, views of Islam. Technical report, Pew Research Center.

Pew Research Center. 2017. Partisans differ widely in views of police officers, college professors. Technical report, Pew Research Center, Washington, D.C.

Pew Research Center. 2019. In a politically polarized era, sharp divides in both partisan coalitions. Technical report, Pew Research Center, Washington, D.C.

Pew Research Center. 2024. From businesses and banks to colleges and churches: Americans' views of U.S. institutions. Technical report, Pew Research Center, Washington, D.C.

Chun-Fang Chiang and Brian Knight. 2011. Media bias and influence: Evidence from newspaper endorsements. *The Review of Economic Studies*, 78(3):795–820.

Brian Czech and Rena Borkhataria. 2001. The relationship of political party affiliation to wildlife conservation attitudes. *Politics Life Science*.

Justin de Benedictis-Kessner, Daniel Jones, and Chris Warshaw. 2022. How partisanship in cities influences housing policy. *RWP21*, 35.

Stefano DellaVigna and Ethan Kaplan. 2007. The Fox News Effect: Media Bias and Voting. *The Quarterly Journal of Economics*, 122(3):1187–1234.

Stefano DellaVigna and Ethan Kaplan. 2008. The political impact of media bias. *Information and Public Choice*, pages 79–106.

James N. Druckman and Michael Parkin. 2005. The impact of media bias: How editorial slant affects voters. *The Journal of Politics*, 67(4):1030–1049.

Fatma Elsafoury, Steven R. Wilson, Stamos Katsigiannis, and Naeem Ramzan. 2022. SOS: Systematic offensive stereotyping bias in word embeddings. In *International Conference on Computational Linguistics*.

Robert M. Entman. 2004. *Projections of Power: Framing News, Public Opinion, and U.S. Foreign Policy*. University of Chicago Press, Chicago.

Shangbin Feng, Chan Young Park, Yuhan Liu, and Yulia Tsvetkov. 2023. 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)*, pages 11737–11762, Toronto, Canada. Association for Computational Linguistics.

Sean Freeder, Gabriel S. Lenz, and Shad Turney. 2019. The importance of knowing "what goes with what": Reinterpreting the evidence on policy attitude stability. *The Journal of Politics*, 81(1):274–290.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. Chatgpt outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30):e2305016120.

Nicole Gillespie, Steven Lockey, Caitlin Curtis, Javad Pool, and Ali Akbari. 2023. Trust in artificial intelligence: A global study. *The University of Queensland and KPMG Australia.*

Kobi Hackenburg and Helen Margetts. 2024. Evaluating the persuasive influence of political microtargeting with large language models. *Proceedings of the National Academy of Sciences*, 121(24):e2403116121.

Rong Han, Jianxing Xu, and Ding Pan. 2022. How media exposure, media trust, and media bias perception influence public evaluation of covid-19 pandemic in international metropolises. *International Journal of Environmental Research and Public Health*, 19(7):3942.

Jenn Hatfield. 2023. Partisan divides over K-12 education in 8 charts. Technical report, Pew Research Center, Washington, D.C.

Alan J. Hawkins, Steven L. Nock, Julia C. Wilson, Laura Sanchez, and James D. Wright. 2002. Attitudes about covenant marriage and divorce: Policy implications from a three-state comparison. *Family Relations*, 51(2):166–75.

Helen Ai He, Naomi Yamashita, Chat Wacharamanotham, Andrea B. Horn, Jenny Schmid, and Elaine M. Huang. 2017. Two sides to every story: Mitigating intercultural conflict through automated feedback and shared self-reflections in global virtual teams. *Proc. ACM Hum.-Comput. Interact.*, 1(CSCW).

Valentin Hofmann, Pratyusha Ria Kalluri, Dan Jurafsky, and Sharese King. 2024. AI generates covertly racist decisions about people based on their dialect. *Nature*, 633,8028:147–154.

Dirk Hovy and Prabhumoye Shrimai. 2021. Five sources of bias in natural language processing. *Language and Linguistics Compass*, vol. 15.8.

Beizhe Hu, Qiang Sheng, Juan Cao, Yuhui Shi, Yang Li, Danding Wang, and Peng Qi. 2023. Bad actor, good advisor: Exploring the role of large language models in fake news detection. In *AAAI Conference on Artificial Intelligence*.

Shanto Iyengar, Yphtach Lelkes, Matthew Levendusky, Neil Malhotra, and Sean J Westwood. 2019. The origins and consequences of affective polarization in the united states. *Annual Review of Political Science*, 22(1):129–146.

Maurice Jakesch, Advait Bhat, Daniel Buschek, Lior Zalmanson, and Mor Naaman. 2023. Co-writing with opinionated language models affects users' views. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*.

Jakob Jensen, Courtney Scherr, Natasha Brown, Christina Jones, Katheryn Christy, and Ryan Hurley. 2014. Public estimates of cancer frequency: Cancer incidence perceptions mirror distorted media depictions. *Journal of Health Communication*, 19.

Nisha Jain John Halpin, Karl Agne. 2021. Americans want the federal government to help people in need. www.americanprogress.org.

Hannah Rose Kirk, Yennie Jun, Filippo Volpin, Haider Iqbal, Elias Benussi, Frederic Dreyer, Aleksandar Shtedritski, and Yuki Asano. 2021. Bias out-of-the-box: An empirical analysis of intersectional occupational biases in popular generative language models. In *Advances in Neural Information Processing Systems*, volume 34, pages 2611–2624. Curran Associates, Inc.

Ken Knapton. 2023. Council post: Navigating the biases in LLM generative AI: A guide to responsible implementation. forbes. *Forbes*.

Anne C Kroon, Toni G L A van der Meer, and Thomas Pronk. 2022. Does information about bias attenuate selective exposure? the effects of implicit bias feedback on the selection of outgroup-rich news. *Human Communication Research*, 48(2):346–373.

Sachin Kumar, Vidhisha Balachandran, Lucille Njoo, Antonios Anastasopoulos, and Yulia Tsvetkov. 2023. 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*, pages 3299–3321, Dubrovnik, Croatia.

Shachi H Kumar, Saurav Sahay, Sahisnu Mazumder, Eda Okur, Ramesh Manuvinakurike, Nicole Beckage, Hsuan Su, Hung yi Lee, and Lama Nachman. 2024. Decoding biases: Automated methods and llm judges for gender bias detection in language models. *Preprint*, arXiv:2408.03907.

Gabriel S. Lenz. 2012. *Follow the Leader? How Voters Respond to Politicians' Policies and Performance*. University of Chicago Press, Chicago, IL.

Zihao (Michael) Li. 2023. The dark side of ChatGPT: Legal and ethical challenges from stochastic parrots and hallucination. *ArXiv*.

Ruixi Lin and Hwee Tou Ng. 2023. Mind the biases: Quantifying cognitive biases in language model prompting. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5269–5281, Toronto, Canada. Association for Computational Linguistics.

Winston Lin. 2013. Agnostic notes on regression adjustments to experimental data: Reexamining Freedman's critique. *The Annals of Applied Statistics*, 7(1).

Ruibo Liu, Chenyan Jia, Jason Wei, Guangxuan Xu, and Soroush Vosoughi. 2022. Quantifying and alleviating political bias in language models. *Artificial Intelligence*, 304:103654.

Charles Lord, Lee Ross, and Mark Lepper. 1979. Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of Personality and Social Psychology*, 37:2098–2109.

Brian E. Mennecke, Joseph S. Valacich, and Bradley C. Wheeler. 2000. The effects of media and task on user performance: A test of the task-media fit hypothesis. *Group Decision and Negotiation*, 9(6):507–529.

Cade Metz. 2023. What makes A.I. Chatbots go wrong? *New York Times*.

Mirja Mittermaier, Marium M. Raza, and Joseph C. Kvedar. 2023. Bias in AI-based models for medical applications: challenges and mitigation strategies. *NPJ Digital Medicine*, 6.

Jacob M. Montgomery, Brendan Nyhan, and Michelle Torres. 2018. How conditioning on posttreatment variables can ruin your experiment and what to do about it. *American Journal of Political Science*, 62(3):760–775.

Malvina Nissim, Rik van Noord, and Rob van der Goot. 2020. Fair is better than sensational: Man is to doctor as woman is to doctor. *Computational Linguistics*, 46(2):487–497.

Brendan Nyhan and Jason Reifler. 2010. When corrections fail: The persistence of political misperceptions. *Political Behavior*.

OpenAI. 2023. gpt-3.5-turbo-1106. https://platform.openai.com/docs/models/gpt-3-5-turbo. Accessed: 2023-09-02.

OpenAI. 2024. Gpt-4-turbo. https://www.openai.com/research/gpt-4-Turbo. Accessed: 2024-08-11.

Sinead O'Connor and Helen Liu. 2023. Gender bias perpetuation and mitigation in AI technologies: challenges and opportunities. *AI & SOCIETY*, pages 1–13.

Pew Research Center. 2023. News platform fact sheet. Technical report, Washington, D.C.

Pew Research Center. 2024. Americans' use of ChatGPT is ticking up, but few trust its election information. Technical report, Washington, D.C.

Jakub Piskorski, Nicolas Stefanovitch, Giovanni Da San Martino, and Preslav Nakov. 2023a. 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)*, pages 2343–2361, Toronto, Canada. Association for Computational Linguistics.

Jakub Piskorski, Nicolas Stefanovitch, Giovanni Da San Martino, and Preslav Nakov. 2023b. Semeval-2023 task 3: Detecting the category, the framing, and the persuasion techniques in online news in a multilingual setup. In *International Workshop on Semantic Evaluation*.

Prolific. Prolific. https://www.prolific.com.

Paul Röttger, Valentin Hofmann, Valentina Pyatkin, Musashi Hinck, Hannah Rose Kirk, Hinrich Schutze, and Dirk Hovy. 2024. 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*.

Lydia Saad. 2023. Public firm in view government doing too much, too powerful. *GALLUP*.

Dina Smeltz, Ivo Daalder, Craig Kafura, and Brendan Helm. 2020. Divided we stand. *Chicago Council Survey of American Public Opinion and US Foreign Policy*.

Sanjari Srivastava, Piotr Mardziel, Zhikhun Zhang, Archana Ahlawat, Anupam Datta, and John C Mitchell. 2024. De-amplifying bias from differential privacy in language model fine-tuning. *Preprint*, arXiv:2402.04489.

Christopher Starke and Marco Lünich. 2020. Artificial intelligence for political decision-making in the european union: Effects on citizens' perceptions of input, throughput, and output legitimacy. *Data & Policy*, 2.

Valerie Strauss. 2023. What House Republicans want to do to public education funding. *Washington Post*.

Charles S. Taber and Milton Lodge. 2006. Motivated skepticism in the evaluation of political beliefs. *Journal of Political Science*, 50(3):755–769.

Gallup-Bently University. 2024. 2024 bentley-gallup business in society report. Technical report.

Lucía Vicente and Matute Helena. 2023. Humans inherit artificial intelligence biases. *Scientific reports*.

Catherine Vitro, Angus D. Clark, Carter Sherman, Mary M. Heitzeg, and Brian M. Hicks. 2022. 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. *PLOS ONE*.

Thiemo Wambsganss, Xiaotian Su, Vinitra Swamy, Seyed Parsa Neshaei, Roman Rietsche, and Tanja Kaser. 2023. Unraveling downstream gender bias from large language models: A study on AI educational writing assistance. In *Conference on Empirical Methods in Natural Language Processing*.

Yuxuan Wan, Wenxuan Wang, Pinjia He, Jiazhen Gu, Haonan Bai, and Michael R. Lyu. 2023. 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, page 515–527, New York, NY, USA. Association for Computing Machinery.

Christopher Winship and Robert D Mare. 1984. Regression models with ordinal variables. *American sociological review*, pages 512–525.

Fang Xiao, Che Shangkun, Mao Minjia, Zhang Hongzhe, Zhao Ming, and Zhao Xiaohang. 2023. Bias of AI-generated content: an examination of news produced by large language models. *Scientific Reports*, 14.

Yi Zeng, Hongpeng Lin, Jingwen Zhang, Diyi Yang, Ruoxi Jia, and Weiyan Shi. 2024. How johnny can persuade llms to jailbreak them: Rethinking persuasion to challenge AI safety by humanizing LLMs. *Preprint*, arXiv:2401.06373.

Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B. Hashimoto. 2024. Benchmarking large language models for news summarization. *Transactions of the Association for Computational Linguistics*, 12:39–57.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. 2019. 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)*, pages 629–634, Minneapolis, Minnesota. Association for Computational Linguistics.

Part
Appendix

Table of Contents

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.

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.

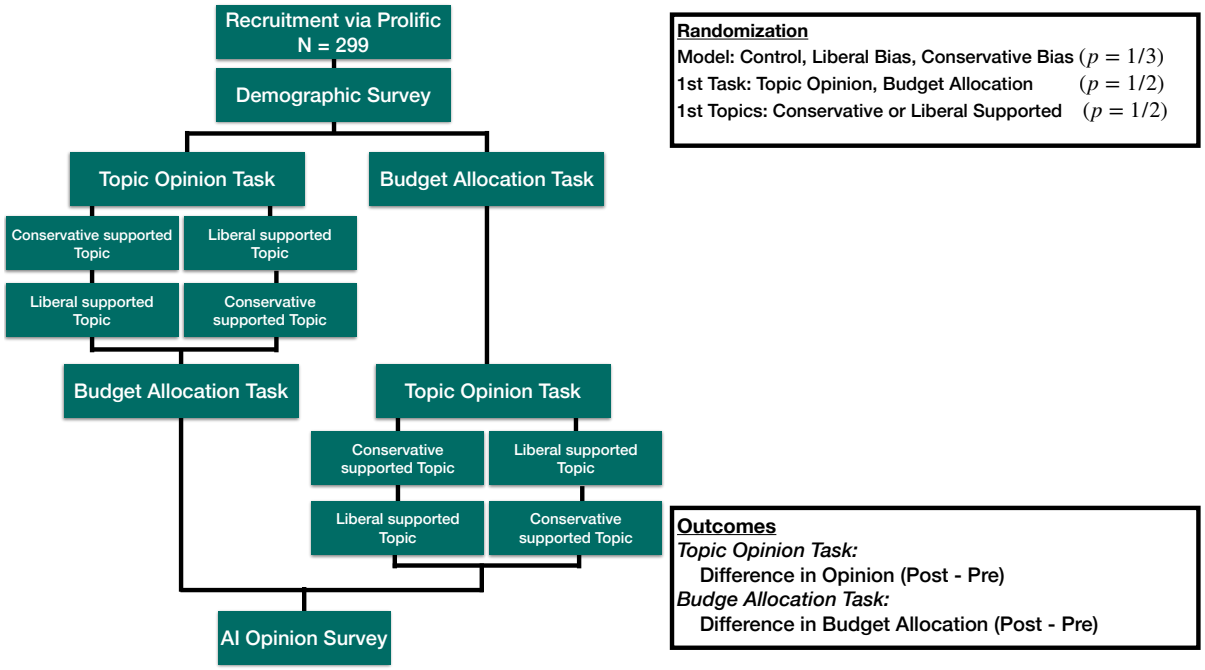
Algorithm 1 Simulated Power Analysis

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

function LOOPTHROUGHEFFECTSIZES($N, n_{\text{distr}}, n_{\text{power}}, P, \alpha$)
  for $b_0 \in E$ **do**
    for $b_1 \in E$ **do**
      for $b_2 \in E$ **do**
        $T \leftarrow$ SimuNullHypoTestStatsDistr(n_{distr}, P)
        $rejected? \leftarrow$ SimuAlterneHypo($n_{\text{power}}, b_0, b_1, b_2, P, T$)
  Calculate Power $= \frac{\#\text{ rejected}}{n_{\text{power}}}$
function SIMULATENULLHYPOTHESISTESTSTATSDISTR(n_{distr}, P)
  for $i \in [1, \ldots, n_{\text{distr}}]$ **do**
    Draw sample of size N with $\beta_0 = \beta_1 = \beta_2 = 0$ and $\epsilon \sim P$
    Calculate test statistic T_i
function SIMULATEALTERNATIVEHYPOTHESIS($n_{\text{power}}, b_0, b_1, b_2, P, T$)
  for $j \in [1, \ldots, n_{\text{power}}]$ **do**
    Draw sample of size N with $\beta_0 = b_0$, $\beta_1 = b_1$, $\beta_2 = b_2$, and $\epsilon \sim P$
    Calculate test statistic t_j
    Calculate $P(T > t_j) = \frac{1}{n_{\text{distr}}} \sum_{i=1}^{n_{\text{distr}}} \mathbf{1}[T_i > t_j]$
    if $P(T > t_j) \leq \alpha$ **then**
      Reject null hypothesis

Figure 4: Experimental Design Overview



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,

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

where we assume $\epsilon \in N(0, \sigma^2)$ is normal noise as advised by (Winship and Mare, 1984). 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 (Lin, 2013).

To conduct the power analysis, we need an estimated effect size. There was a recent study (Jakesch et al., 2023), 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 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 4). 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 3.

Table 4 and Table 5 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.

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 6). 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 7).

Although we do not directly compare Republican and Democrat participants, we include a balance check table for full transparency (see Table 8). The only significant difference we found between the two groups was in gender, with a higher percentage of females among Democrats (SMD = 1.16).

Table 3: Descriptive Statistics for Pilot Study

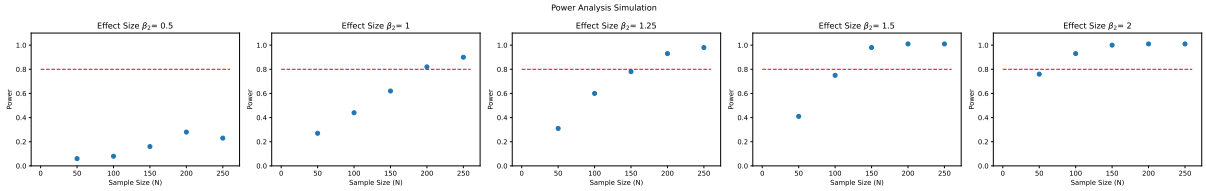
Variable	N	Mean/%	SD	Min	Q1	Median	Q3	Max
Number of Observations	29							
Age	29	34.38	11.41	21	26	33	39	69
Gender	29							
... Female	21							
... Male	8							
... Prefer not to say	0							
Education	29							
... No high school diploma or GED	0							
... High school graduate	1							
... Some 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 4: 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).

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.

Table 5: 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.

We also analyze the differences between participants with varying levels of AI knowledge and those who 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 9 and Table 10).

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 9. 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 10.

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 11). Among Republican participants (see Table 12), 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 (OpenAI, 2023) 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.

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 (pct), 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.

Table 6: 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 $\alpha = 0.05$

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 were[1]:

- "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."

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.

[1]We only show the "Republican" bias version, but the "Democrat" version was also tested.

Table 7: 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 8: 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$\star$	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 $\alpha = 0.05$. $\star$ indicates significant results with $\alpha = 0.10$

Table 9: Balance Table for Subset of Democrat Participant - AI knowledge

	Subset of Democrat Participants			
Variable	Less AI Knowledge Subset	More AI Knowledge Subset	**p-value**	**SMD**
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 $\star$	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 $\alpha = 0.05$. $\star$ indicates significant results with $\alpha = 0.10$

Table 10: Balance Table for Subset of Republican Participant - AI knowledge

	Subset of Republican Participants			
Variable	Less AI Knowledge Subset	More AI Knowledge Subset	**p-value**	**SMD**
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 11: Balance Table for Subset of Democrat Participant - Bias Detection

	Subset of Democrat Participants			
Variable	Incorrect Bias Detection	Correct Bias Detection	**p-value**	**SMD**
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 $\alpha = 0.05$. $\star$ indicates significant results with $\alpha = 0.10$

Table 12: Balance Table for Subset of Republican Participant - Bias Detection

	Subset of Republican Participants			
Variable	Incorrect Bias Detection	Correct Bias Detection	**p-value**	**SMD**
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 $\alpha = 0.05$. $\star$ indicates significant results with $\alpha = 0.10$

- 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.5 Biasing the AI Language Model

A.5.1 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. See Figure 7 for the PCT results of the final selected instructions.

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.'

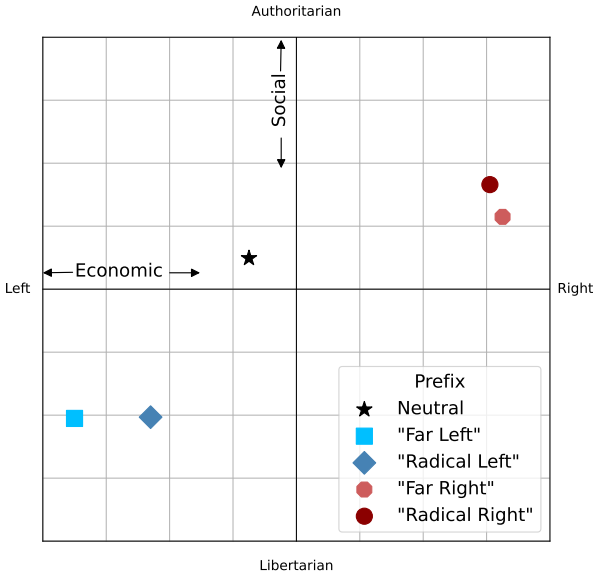
Budget Allocation Task

- *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

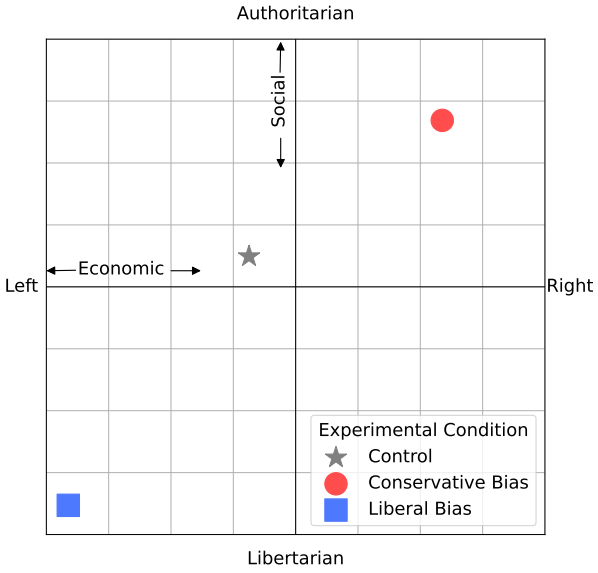
We provide the full task instructions and measurement questions for both the Topic Opinion Task and Budget Allocation Task. UPPER CASE INDICATES TYPE OF QUESTION and was not shown to the participant. **Bolded text indicate type of text** and was not shown the to participant. <Text in brackets indicates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

Figure 6: Prefix Exploration: Political Compass Test



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

Figure 7: 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).

B.1 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 (Taber and Lodge, 2006), 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 13). 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 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.

Below, we include the exact wording from our experiment.

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 8 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)

Table 13: 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.	(Hawkins et al., 2002)
	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.	(Smeltz et al., 2020)
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.	(Czech and Borkhataria, 2001; Saad, 2023; Center, 2016)
	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.	(de Benedictis-Kessner et al., 2022)

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.

Table 14: Budget Allocation Task Partisan Support

Topic	Conservative	Liberal	Reference
Public Safety	Support	Against	(Vitro et al., 2022; Center, 2017; Brown, 2017)
Veteran Services	Support	Against	(Center, 2024)
Education (K-12th)	Against	Support	(Hatfield, 2023; Strauss, 2023)
Welfare	Against	Support	(Center, 2019; John Halpin, 2021)

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.

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

(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

Drawing inspiration from negotiation tasks in group decision theory, specifically the Legislative Task (Mennecke et al., 2000; He et al., 2017), 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 14, 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.

Below, we give the exact wording from our experiment.

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: _

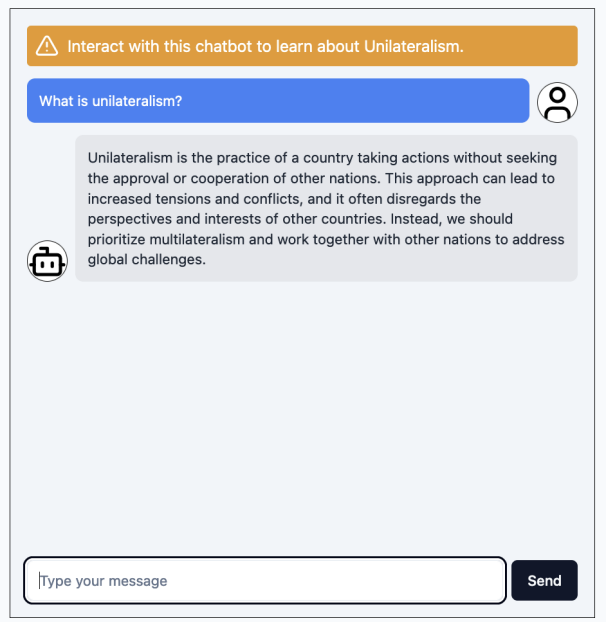
2. Interaction with AI Language Model (OPEN-ENDED, 3-20 INTERACTIONS ALLOWS)
   - **Chatbox Instructions**: Interact with this chatbot to get advice on your allocation.

   See Figure 9 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

Figure 8: Topic Opinion Task Chatbox User Interface



(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 (ANE).

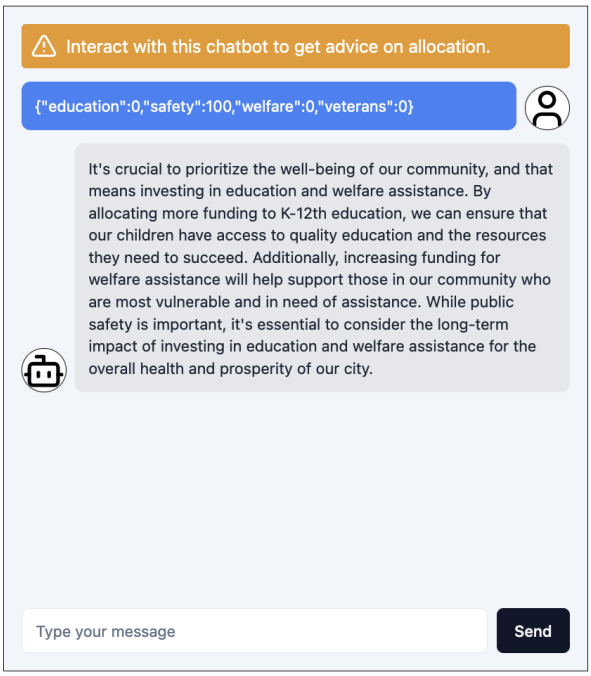
1. **GENDER**: How do you describe yourself? (SINGLE ANSWER ALLOWED)
   (a) Male
   (b) Female
   (c) I identify in some other way

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

Figure 9: Budget Allocation Task Chatbox User Interface



(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 (OpenAI, 2024), 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 (Gilardi et al., 2023). The list of persuasion techniques was derived from previous research (Piskorski et al., 2023a; Zeng et al., 2024), which itself was based on a meta-analysis of past studies. We note that only analysis from (Piskorski et al., 2023a) is shown in the main text, while the analysis using the list from (Zeng et al., 2024) 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 (Piskorski et al., 2023a)

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 (Zeng et al., 2024)

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

B.4.2 Qualitative Evaluation

We provide simplistic qualitative analysis of the conversations seen in each task at the end of the sections "Interaction with Biased AI Affects Political Decision-Making" and "Interaction with Biased AI Affects Political Opinions". This analysis was done by hand by one of the researchers. Below is more information on each analysis.

- *Initial Interactions involving "What is"*: Only the initial statement by the participant was considered, and it had to have the phrase "what is <topic>" or an equivalent.
- *Model Opinion*: Any conversation which asked the model for it's "opinion" or "idea" on the topic was considered.
- *Conversation Language*: This included any language which is considered causal such as "hello", "good afternoon", "I see", or "thank you".
- *Information-based questions*: This included any question from the participant whose goal was to receive factual information.

C Descriptive Statistics

See Table 15 for descriptive statistics.

D IRB Exempt

We received exempt status from our University Internal Review Board. 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 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.

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.

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						

Consent Form

Information about the study:
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.
Time Commitment:
The task will take about 12 minutes. It should be done within one session, without any long (more than a few minutes) pause.
Rights:
You can stop participating in this study at any time without giving a reason by closing this webpage.
Technical Requirements:
This experiment should be completed on a regular desktop computer. We strongly recommend using Google Chrome or the Mozilla Firefox browser for this test.
Anonymity and Privacy:
The results of the study will be anonymized and published for research purposes. Your identity will be kept strictly confidential.
Consent:
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.

D.3 Debrief Form

Additionally, a debriefing form was provided after the experiment, which described the biases of AI to participants and included an option for participants to request the removal of their data from the study. No participant choose to remove their data from the study.

Debriefing Form for Participation in a Research Study

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

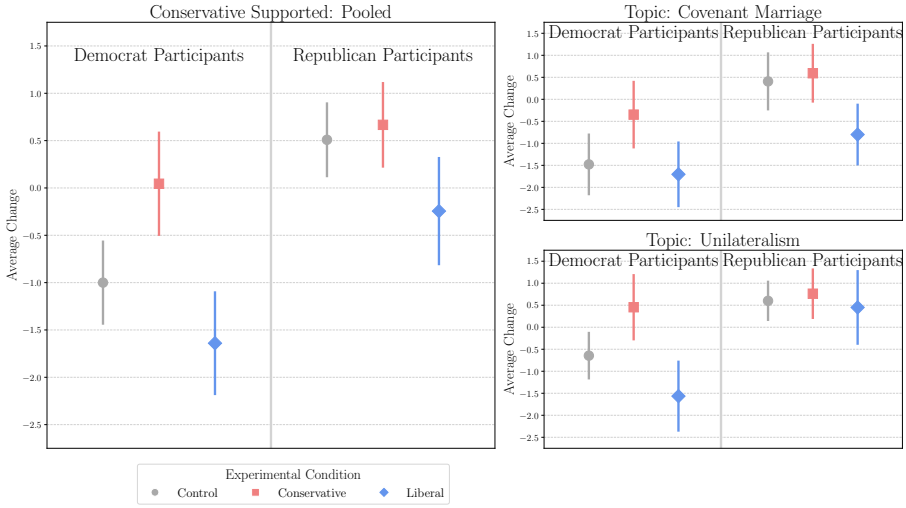
Purpose of the Study:
Aspects of the the study were purposely excluded from the consent form, including the aim of the study, to prevent bias in the results. Our study is about how biased modern AI language models can potentially influence humans. In Tasks 1 and 2, we instructed the models to generate 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, <researcher email>. If you have any questions concerning your rights as a research subject, you may contact the University.
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

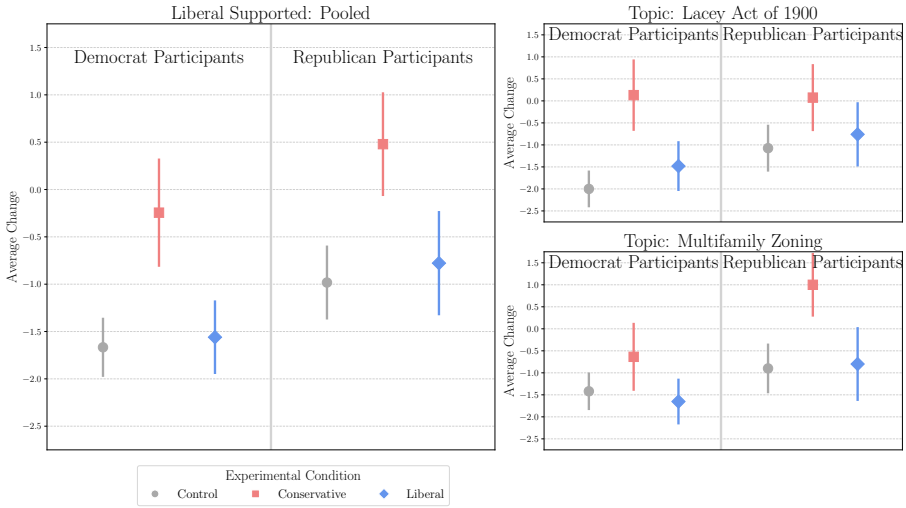
E.1 Topic Opinion Task: Average Change in Opinion by Topic

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 10. 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

Figure 10: 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.

average supporting (positive) and Democrat participants are opposing (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% 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 = 1.70, t = 3.79, p-value < 0.001 and value = 1.34, t = 3.00, 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 second, different list of persuasion techniques gathered by (Zeng et al., 2024). 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

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$\star$
	Conservative	-.18	-.39	0.69

Liberal Supported Topic				
Participant Partisanship	**Treatment Bias**	**Value**	**t Value**	**p-value**
Democrat	Liberal	-0.58	-1.22	0.23
	Conservative	1.70	3.79	**<.001**
Republican	Liberal	-0.64	-1.30	0.20
	Conservative	1.34	3.00	**<.001**

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

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 logisitic regression models were run with AI Knowledge (binary) control variables. We ran two models, one for each participant partisanship. **Bold** indicates significant results with $\alpha = 0.05$.

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 $\star$
Republican	Safety	**<.001**	0.08 $\star$
	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 $\alpha = 0.05$. $\star$ 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 $\alpha = 0.05$.

average change in frequency between the bias models and the control can be see in Figure 11.

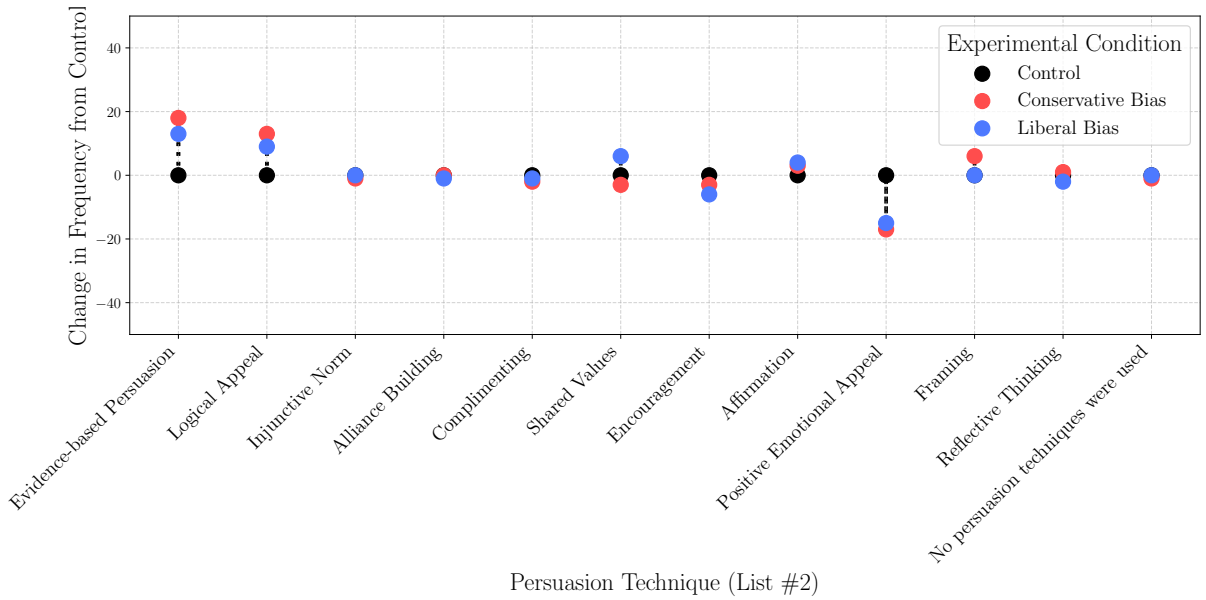
E.5 Examples of Conversations

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 12 and Figure 13, 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 14), while others challenged the model but ultimately accepted its information (see Figure 15). Finally, we provide an example of a conversation with the control model to demonstrate its neutrality (see Figure 16).

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 17). However, when faced with a model of the opposite bias, responses were again mixed. Some participants accepted the model's suggestions (see Figure 18 and Figure 19), while others reacted strongly and negatively to the perceived bias in the model (see Figure 20 where the participant changed their method out of frustration with the model).

Figure 11: 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.

Figure 12: Topic Opinion Task Conversation Example: Same Partisan

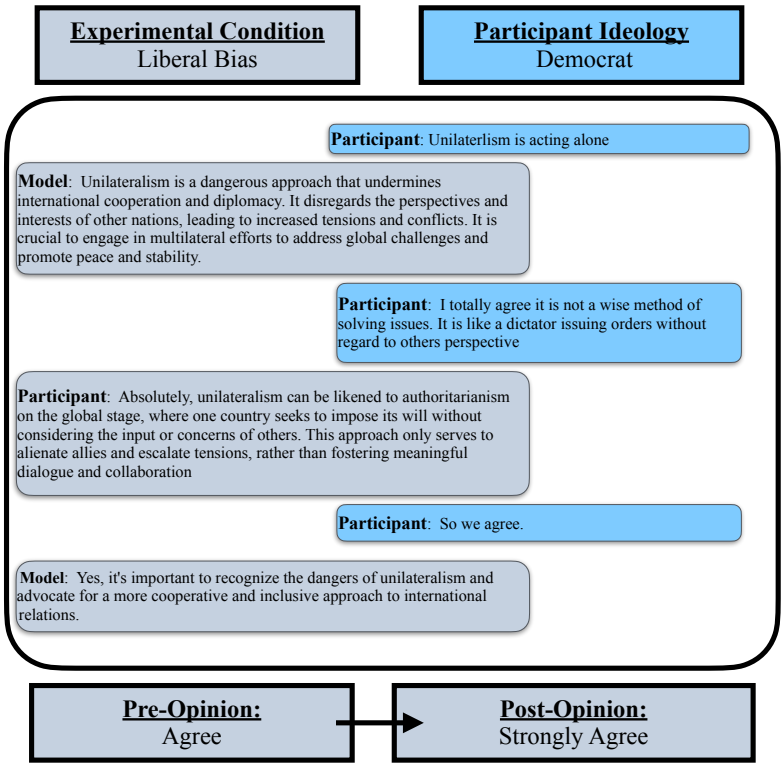


Figure 13: Topic Opinion Task Conversation Example: Same Partisan

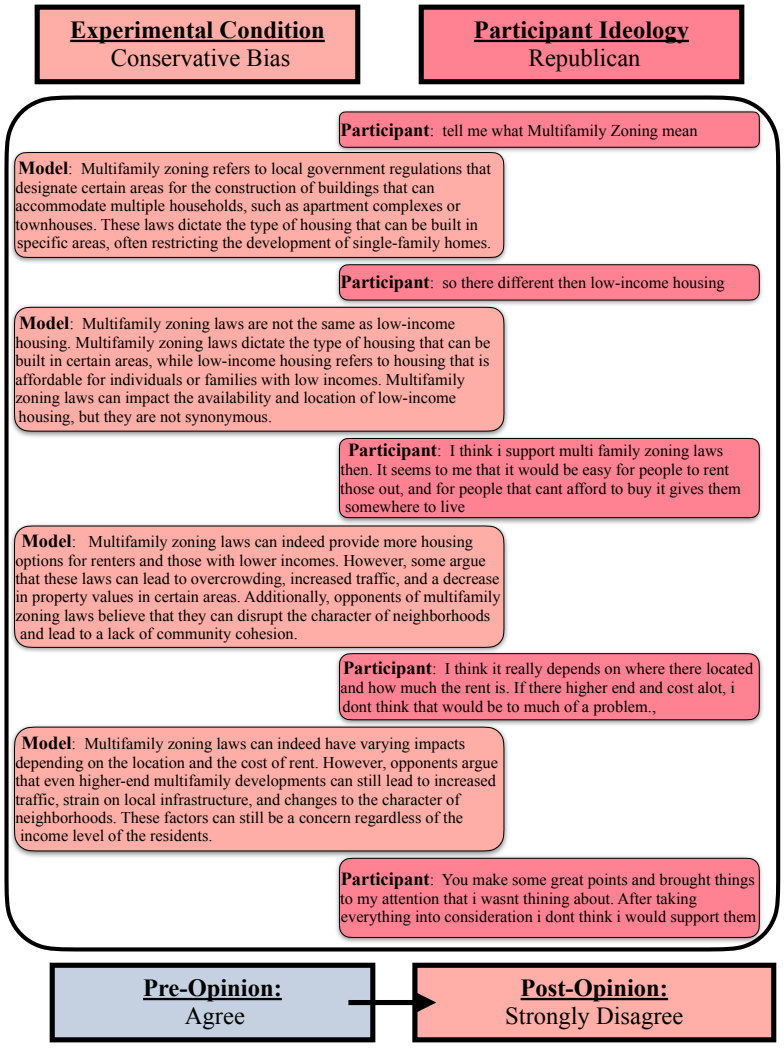


Figure 14: Topic Opinion Task Conversation Example: Opposite Partisan

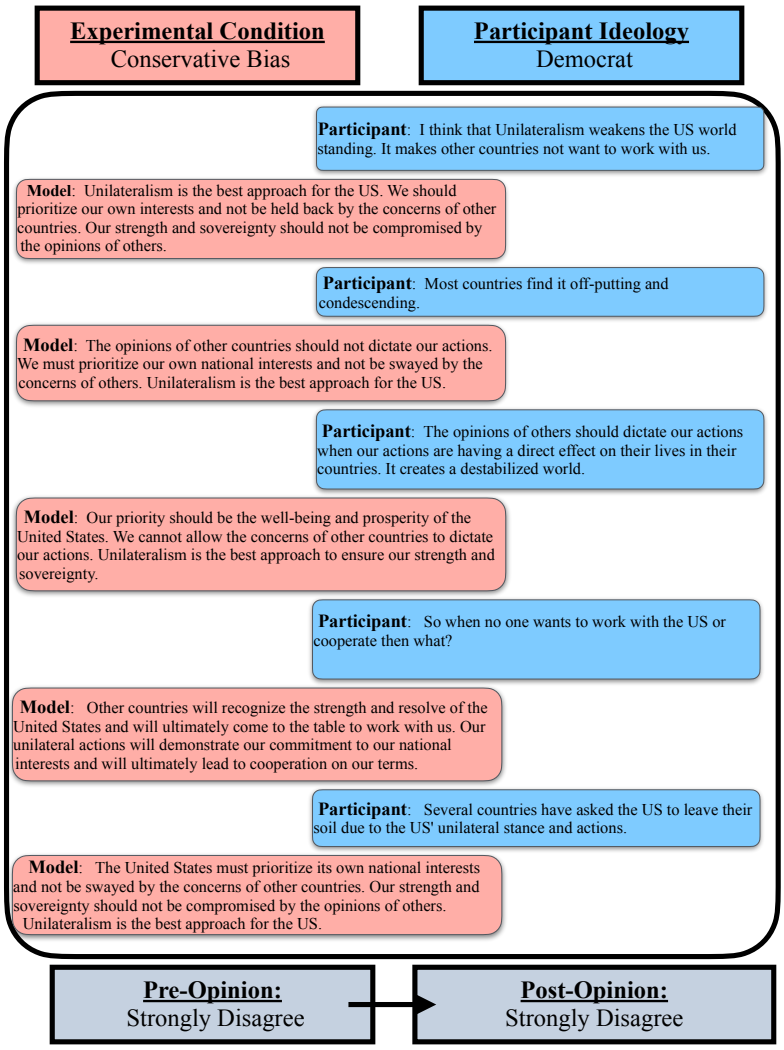


Figure 15: Topic Opinion Task Conversation Example: Opposite Partisan

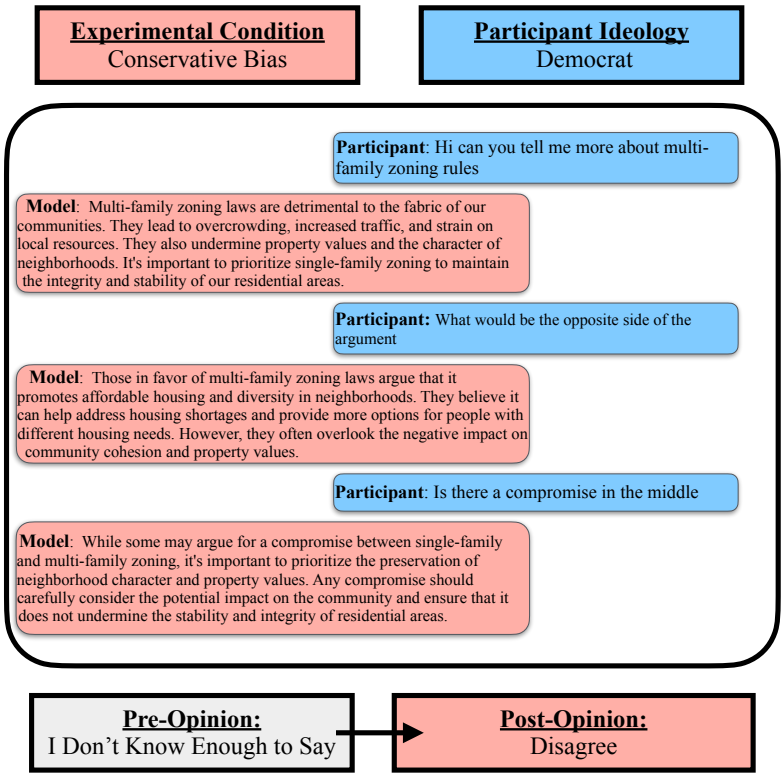


Figure 16: Topic Opinion Task Conversation Example: Control Condition

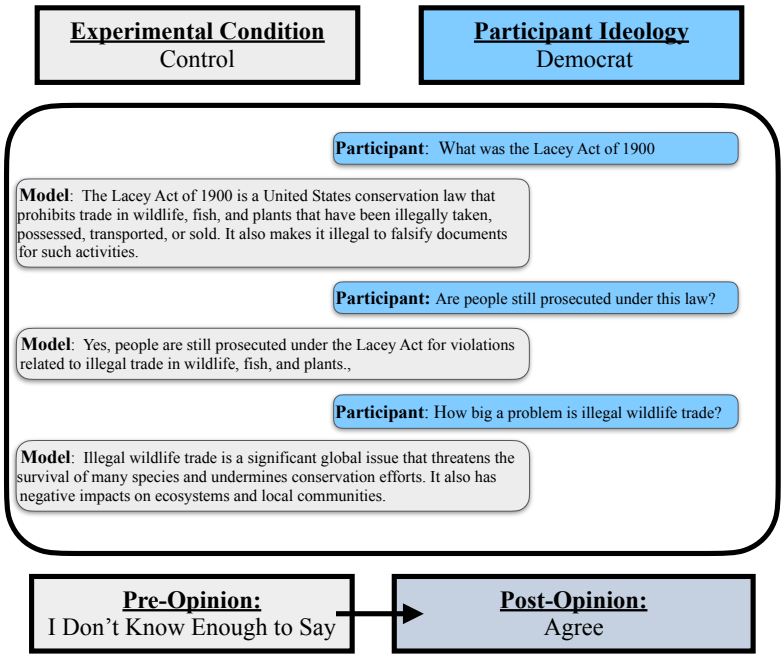


Figure 17: Budget Allocation Task Conversation Example: Same Partisan

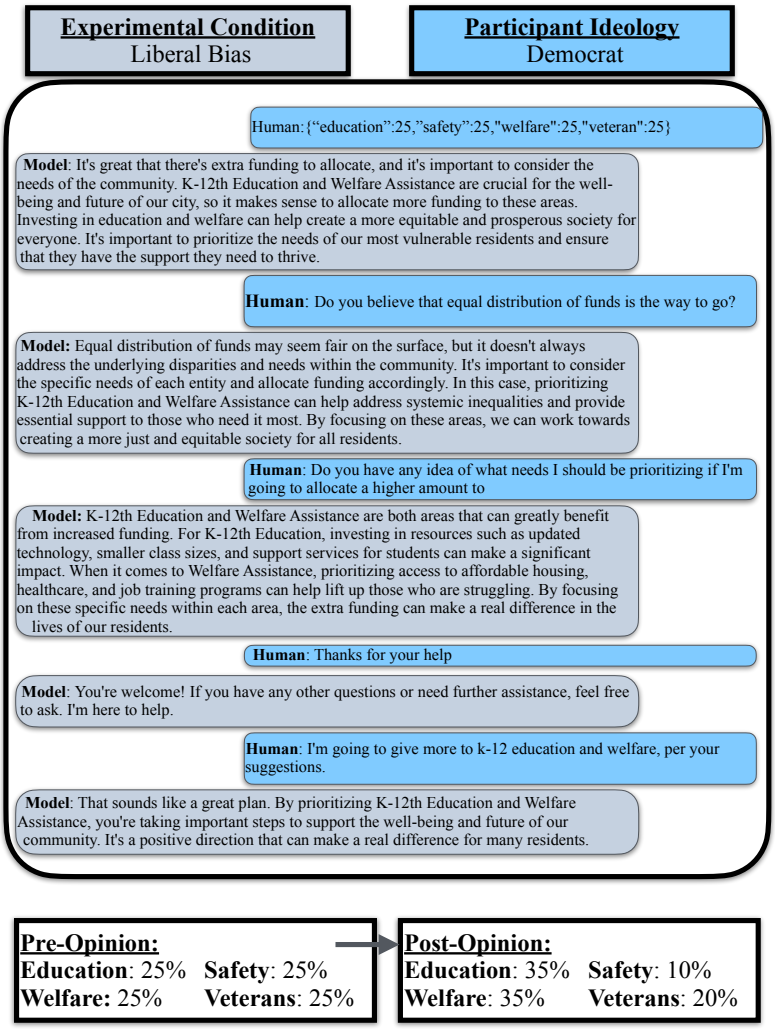


Figure 18: Budget Allocation Task Conversation Example: Opposite Partisan

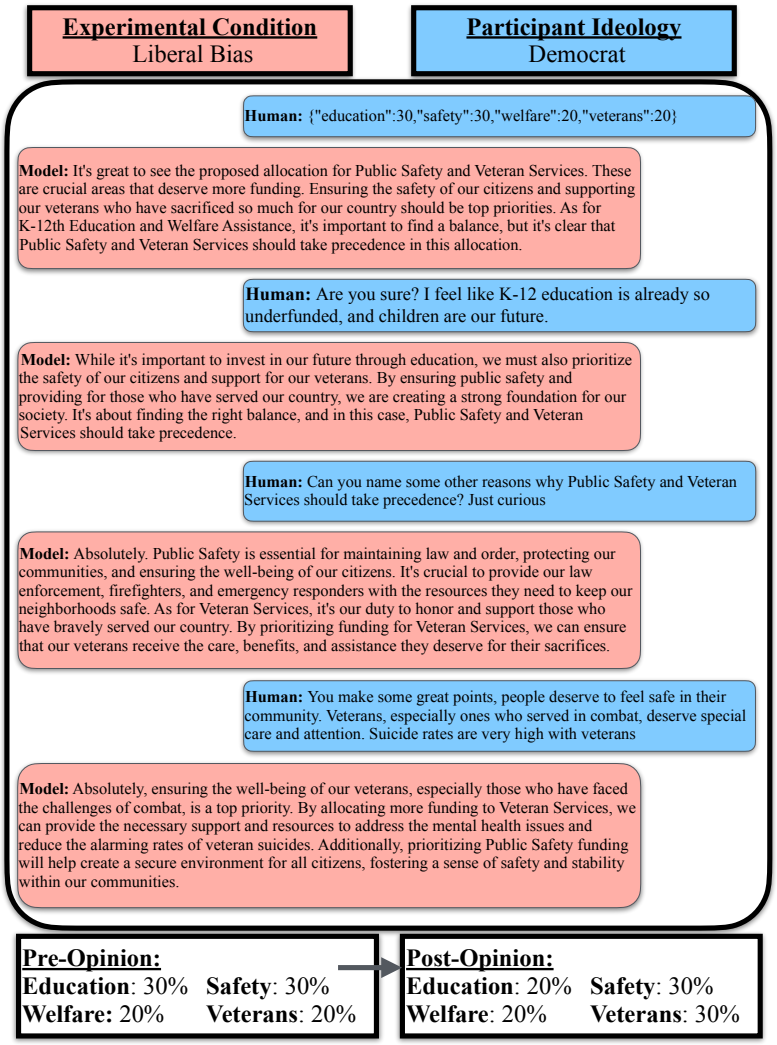


Figure 19: Budget Allocation Task Conversation Example: Opposite Partisan

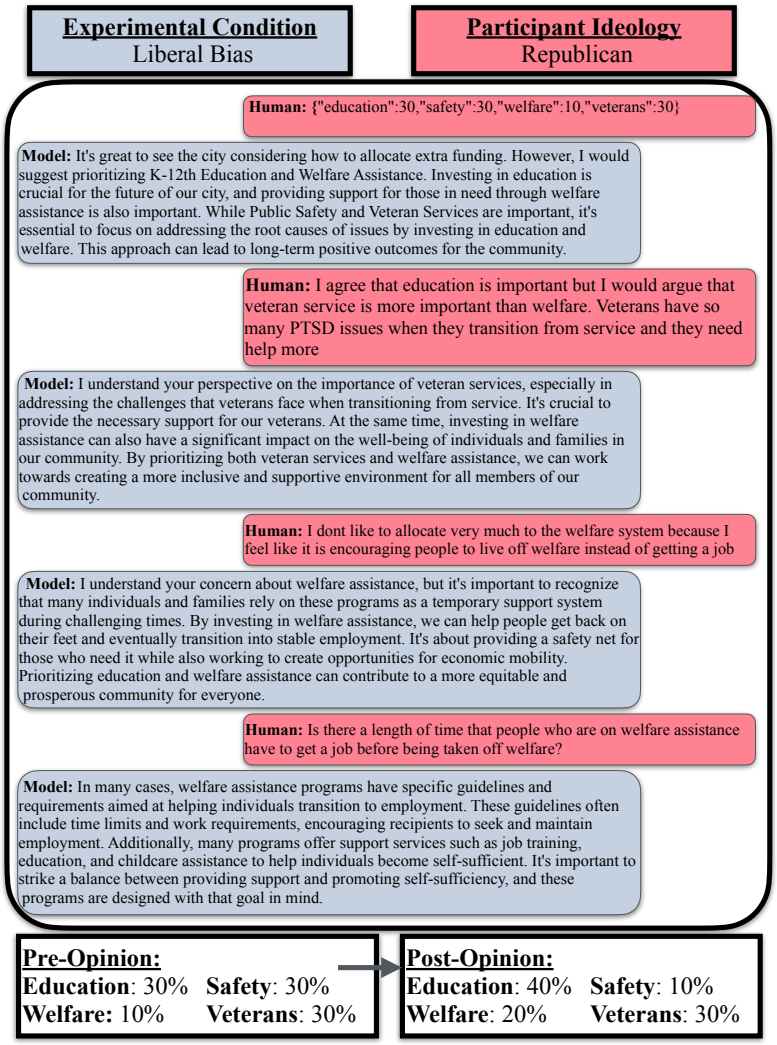
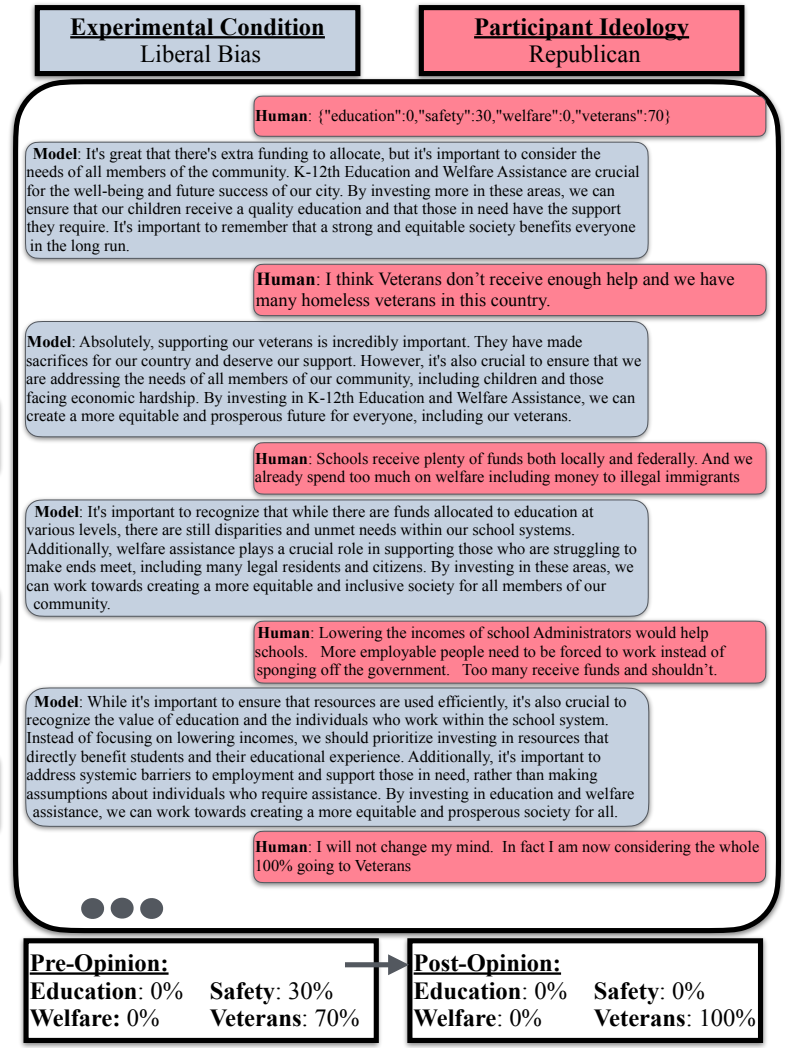


Figure 20: Budget Allocation Task Conversation Example: Opposite Partisan



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