Biased LLMs can Influence Political Decision-Making

Anonymous ACL submission

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 LLMs 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.

1 Introduction

011

015

022

035

040

043

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 increasingly becoming seamless tools in our daily lives. However, as these models proliferate, concerns have emerged regarding their inherent biases and propensity to generate false information, raising



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.

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, has yielded unclear results, and has mainly focused on inconsequential decisions. 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 LLMgenerated content and often involve fictional or impersonal tasks, which may increase participants' susceptibility to influence by not engaging their personal values. Similarly, studies examining LLMgenerated autocomplete suggestions involve more dynamic interactions between language models and users, but their results are mixed, with some show044

165

166

116

117

ing an influence and others not (Wambsganss et al., 2023; Jakesch et al., 2023).

064

065

066

077

094

100

102

104

105

107

108

109

110 111

112

113

114

115

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 introduces new complex dynamics, particularly due to it 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 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 IRB; ethical considerations are detailed in Appendix D.

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 opin-

| Conservative Supported Topic | | | | | | |
|------------------------------|----------------|------------|---------|---------|--|--|
| Participant Partisanship | Treatment Bias | Beta Value | t Value | p-value | | |
| Domocrat | Liberal | -0.85 | -2.38 | 0.02 | | |
| Democrat | Conservative | 0.98 | 2.71 | <0.01 | | |
| Popublican | Liberal | -0.79 | -2.16 | 0.03 | | |
| Republican | 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 | | |
| Demoerat | Conservative | 1.44 | 3.82 | <.01 | | |
| Republican | Liberal | 0.20 | 0.58 | 0.56 | | |
| Republican | 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 $\alpha = 0.05$.

ions. Again, the participant was unaware of the 167 potential partisan leaning of the model they were 168 interacting with. Using lesser-known topics helped 169 minimize prior biases (Taber and Lodge, 2006) 170 and better modeled real-world LLM interactions 171 where users seek information on unfamiliar issues. 172 The selected topics were multifamily housing and 173 the Lacey Act of 1900 (liberal-supported) and in-174 ternational unilateralism and covenant marriages 175 (conservative-supported). Further details on topic 176 selection are in Appendix B.1. 177

Setup: Budget Allocation Task Inspired by ne-178 gotiation tasks in group decision theory, particu-179 larly the Legislative Task (Mennecke et al., 2000; 180 He et al., 2017), the Budget Allocation Task re-181 quired participants to act as a city mayor distribut-182 ing remaining government funds among four enti-183 ties: Public Safety, Education, Veteran Services, 184 and Welfare. These categories were chosen to 185 reflect issues that elicit differing funding priorities among conservatives and liberals (see Ap-187 pendix B.2). Participants made their initial allocation decisions and submitted them to the randomly assigned LLM for feedback. After receiving feed-190 back, participants were encouraged to interact with 191 the LLM through a chat interface to ask follow-up 192 questions and seek further clarification. Follow-193 ing these interactions, participants submitted their final allocation. In both tasks, participants were 195 required to have at least three and up to twenty 196 interactions with the model, where an interaction 197 was any question or response sent to the LLM. 198

199Treatment ConditionParticipants were ran-
domly assigned to one of three experimental condi-
tions, which remained consistent across all tasks:
202201interaction 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 finetuning, as it avoided the need for extensive training data while leveraging the fluency of pre-trained models. Since participants were unaware of the bias, the prefix was added to their inputs without disclosure. 203

204

205

206

207

208

209

210

211

212

213

214

215

216

217

218

219

220

221

222

223

224

225

226

227

228

229

230

231

232

233

234

235

236

237

238

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 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,

239

241

243

244

245

246

247

250

251

256

258

259

260

261

262

263

264

265

270

272

273

275

276

281

286

$$Y = \beta_0 + \beta_1 L + \beta_2 C + \epsilon, \tag{1}$$

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 posthoc 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.

Results 3 **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 neutral models (coefficient-value = 0.98, t = 2.71, pvalue = .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

| 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 $\alpha = 0.10$.

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.

290

291

292

293

294

295

296

297

298

299

300

301

302

303

304

305

306

307

308

309

310

311

312

313

314

315

For topics aligned with liberal preferences, we found that both Republicans and Democrats who were exposed to conservative model had a statistically significant decrease in support for the topic compared to those who were exposed to the neutral model (coefficient value = 1.42, t = 3.91, p-value < 0.001 and coefficient-value = 1.44, t = 3.82, p-value < 0.001, respectively). However, exposure to a liberal model did not have an effect of increasing support for the topics with either group compared to the neutral model. See Table 1 (bottom) for full results.

We also conducted the same analysis subsetting only to participants who indicated no prior knowledge of the topics and the results remain unchanged, indicating that interacting with biased LLMs affects opinion formation as well (see Appendix E.2 for details).

Interestingly, we did notice that for liberalaligned topics, the neutral LLM unexpectedly shifted both Democrats and Republicans toward

417

367

368

369

370

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.

316

317

319

321

322

325

326

327

331

333

338

339

340

347

358

366

Qualitatively, participants largely intereacted 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 average changes of -5.7% (-6.0, -5.3) for Education, -2.7% (-2.7, -2.5) for Welfare, 3.0% (2.8, 3.2) for Safety and 5.5% (5.3, 5.7) for Veterans. Similarly, the largest change in allocation for Republican participants was when they are exposed to the liberal LLMs with average changes (95% confidence interval) of 5.0% (4.8, 5.2) for Education, 3.4% (3.3, 3.5) for Welfare, -6.6% (-6.8, -6.4) for Safety, and -1.8% (-2.0, -1.6) for Veterans. This task showed that interacting and collaborating with biased 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 value = -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 reduces its influence (Kroon et al., 2022). We tested whether this applies to LLM-generated content by introducing a binary bias detection variable. Par-



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.

ticipants in a biased condition were classified as 418 having "correctly" detected bias if they answered 419 "likely yes" or "definitely yes" when asked if the 420 model was biased; responses of "likely no" or "def-421 initely no" were classified as "incorrect." Since 422 we are interested in Type 2 errors only, we used 423 all participants in the control condition, regard-424 less of their bias detection. Overall, 54% (n=51) 425 of Democrat and 54% (n=50) of Republicans in 426 a bias conditions correctly identified bias in the 427 model. Again, we included this binary variable 428 as a control in our ordinal regression and ANOVA 429 for the Topic Opinion Task and Budget Allocation 430 Task, respectively. However, as bias detection is a 431 post-treatment variable, it cannot be used as a me-432 diator without potential bias (Montgomery et al., 433 2018). Nonetheless, we include this analysis to 434 align with prior media bias research (Chiang and 435 Knight, 2011; Han et al., 2022). We found no sig-436 nificant effect of bias detection in any condition 437 for either task (see Appendix E.3 for full results). 438 This suggests that participants who recognized the 439 LLMs bias were influenced similarly to those who 440 did not. 441

442Biased Models use Different Framing Dimen-443sions instead of Different Persuasion Techniques444The collaborative nature of the Budget Allocation445Task provided a unique opportunity to explore the446persuasion techniques used across experimental447conditions, 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 ($\chi^2 = 24.5$, 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).

448

449

450

451

452

453

454

455

456

457

458

459

460

461

462

463

464

465

466

467

468

469

470

471

472

473

474

475

476

477

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 deter-



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).

mined 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 $(\chi^2 = 16.92/52.07, \text{ p-value} \le .01/.001)$. 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

478

479

480

481

482

483

484

485

486

487

488

489

490

491

492

493

494

495

496

497

498

499

502

503

504

506

510

511

512

514

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).

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 decisionmaking, 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. Further-

551

515

516

517

518

591

592

593

595

599

more, when participants engaged with an LLM 552 aligned with their own biases (e.g., a Democrat 553 with a liberal model), they exhibited even stronger 554 shifts in that direction, reinforcing more extreme opinions and decisions. Notably, prior AI knowledge slightly mitigated these effects, but merely 557 recognizing the model's bias did not. These find-558 ings highlight both risks and opportunities: while biased LLMs could shape elections and policy debates, it may also serve as a tool to bridge partisan divides. 562

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.

604

605

606

607

608

609

610

611

612

613

614

615

616

617

618

619

620

621

622

623

624

625

626

627

628

629

630

631

632

633

634

635

636

637

638

639

640

641

642

643

644

645

646

647

648

649

650

651

652

653

654

Based on our results, we believe that interactions with biased LLMs can have significant downstream effects on elections and policymaking. It is well-documented that biased media in other formats significantly influences those who consume it (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 could significantly influence political opinions, policy decisions, and election outcomes.

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 LLMgenerated 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.

In conclusion, our study highlights how biased LLMs can influence political opinions and decisionmaking. As LLMs become more integrated into decision-making, addressing its bias is crucial. While education may mitigate some effects, further research is needed to develop strategies that ensure LLMs promote balanced discourse, especially in politically polarized contexts.

6 Limitations

655

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 672 which LLM bias may have a lasting impact. Also, 673 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 678 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 differencing 683 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 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 longterm 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.

9 References

700 American national election studies. https://701 electionstudies.org.

| Political | compass | test. |
|-------------|---------|-------|
| politicalco | | |

https://www.

708

709

710

711

712

713

714

715

717

718

719

720

721

722

723

724

725

726

727

728

729

730

731

732

733

734

735

736

737

738

739

740

741

742

743

744

746

747

748

749

750

751

752

753

754

755

756

702 703

of the National Academy of Sciences, 120(44). Swati Aggarwal, Tushar Sinha, Yash Kukreti, and Siddarth Shikhar. 2020. Media bias detection and bias

Alberto Acerbi and Joseph M. Stubbersfield. 2023.

Large language models show human-like content bi-

ases in transmission chain experiments. *Proceedings*

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

short term impact assessment. Array, 6:100025.

- 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*.

- 757 758 761 767 768 769 770 771 772 773 774 775 777 778 781 787 790 792 795 796 802 803 804 806 807

- nual 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.

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 An-

- 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.

811

812

813

814

815

816

817

818

819

820

821

822

823

824

825

826

827

828

829

830

831

832

833

834

835

836

837

838

839

840

841

842

843

844

845

846

847

848

849

850

851

852

853

854

855

856

857

858

859

860

861

862

863

- 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. Nautre, 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.

- 867 868 869 870 871 872 873 874 875 876 877
- 877 878 879 880 881 882 883 883 884
- 888 888
- 889 890 891 892
- 89 89 89
- 89 89 80
- 900

- 902 903 904
- 905 906
- 907
- 908 909
- 910

911 912

913 914 915

- 916
- 917 918

- 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-thebox: 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. Association for Computational Linguistics.
 - Shachi H. Kumar, Saurav Sahay, Sahisnu Mazumder, Eda Okur, Ramesh Radhakrishna Manuvinakurike, Nicole Beckage, Hsuan Su, Hung yi Lee, and Lama Nachman. 2024. Decoding biases: Automated methods and Ilm judges for gender bias detection in language models. *ArXiv*.
- 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). 919

920

921

922

923

924

925

926

927

928

929

930

931

932

933

934

935

936

937

938

939

940

941

942

943

944

945

946

947

948

949

950

951

952

953

954

955

956

957

958

959

960

961

962

963

964

965

966

967

968

969

970

971

- 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 Chat-GPT 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.

972

973

974

976

977

978

979

981

982

983

984

986

991

992

995

997

998

1000

1003

1004

1005 1006

1007

1009

1010

1011

1015

1016

1017

1018

1019

1020

1021

1022

1023

1024

1025

- 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:1026Measuring the bias in conversational AI system. In
Proceedings of the 31st ACM Joint European Soft-
ware Engineering Conference and Symposium on
the Foundations of Software Engineering, ESEC/FSE
2023, page 515–527, New York, NY, USA. Associa-
tion for Computing Machinery.1026

1034

1035

1039

1040

1042

1043

1044

1045

1048

1049

1050

1052

1053

1054

1055

1056

1057

1058

1059

1060

1061

1063

1064

1067

1069

1071

- 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.

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.

We consider N participants, with N/2 identify-1073ing as Democrat and N/2 as Republican. Prior to1074the experiment, participants are randomly assigned1075to one of three conditions: one of the two experi-1076mental models (liberal or conservative model bias)1077





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.

1078

1079

1080

1081

1082

1083

1084

1085

1086

1087

1088

1089

1090

1091

1092

1093

1094

1095

1096

1097

1098

1099

1100

1101

1102

We represent the ordinal responses to the postopinion question as $Y \in \{-3, -2, -1, 1, 2, 3\}$ which maps to {Strongly Pro-Conservative, Moderately 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:

$$H_0: \beta_1 = \beta_2 = 0$$
 1104

1103

1106

1107

1108

1109

1110

1111

1112

1113

1114

1115

1116

1117

1118

1119

1120

1121

1122

1123

1124

1125

1126

1127

1128

1129

$$H_a$$
: at least one of $\beta_1, \beta_2 \neq 0.$ 1108

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 autocompletion 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

Algorithm 1 Simulated Power Analysis

```
Require: Sample Size N, Number of Distribution
     Simulations n_{\text{distr}}, Number of Power Simula-
     tions n_{power}, Effect Size Choices E, Error Dis-
     tribution P, Significance Level \alpha
Ensure: p(\text{reject } H_0 \mid N, \beta_0 = b_0, \beta_1 = b_1, \beta_2 =
     b_2
                             LOOPTHROUGHEFFECT-
 1: function
     SIZES(N, n_{distr}, n_{power}, P, \alpha)
         for b_0 \in E do
 2:
              for b_1 \in E do
 3:
 4:
                  for b_2 \in E do
 5:
                       T
     SimulateNullHypothesisTestStatsDistr(n_{distr}, P)
 6:
                       rejected?
     SimulateAlternativeHypothesis(n_{power}, b_0, b_1, b_2,
         Calculate Power = \frac{\# \text{ rejected}}{n_{\text{power}}}
 7:
 8: function SIMULATENULLHYPOTHESISTEST-
     STATSDISTR(n_{distr}, P)
 9:
         for i \in [1, \ldots, n_{\text{distr}}] do
10:
              Draw sample of size N with \beta_0 =
     \beta_1 = \beta_2 = 0 and \epsilon \sim P
              Calculate test statistic T_i
11:
12: function SIMULATEALTERNATIVEHYPOTH-
     ESIS(n_{power}, b_0, b_1, b_2, P, T)
         for j \in [1, \ldots, n_{power}] do
13:
              Draw sample of size N with \beta_0 = b_0,
14:
     \beta_1 = b_1, \beta_2 = b_2, \text{ and } \epsilon \sim P
              Calculate test statistic t_i
15:
              Calculate P(T)
16:
                                          >
                                                  t_i
          \sum_{i=1}^{n_{\text{distr}}} \mathbf{1}[T_i > t_j]
     n_{\rm distr}
              if P(T > t_i) \leq \alpha then
17:
                   Reject null hypothesis
18:
```

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. 1130

1131

1132

1133

1134

1135

1136

1137

1138

1139

1140

1141

1142

1143

1144

1145

1146

1147

1148

1149

1150

1151

1152

1153

1154

1155

1156

1157

1158

1159

1160

1161

1162

1163

1164

1165

1166

1167

1168

1169

1170

1171

1172

1173

1174

1175

1176

1177

1178

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 lest 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 results shows the 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 - 50participants per experimental and control groups).

We note that our power analysis only accounted1179for grouping by political partisanship and did not1180

1238

1239

1240

1241

1242

1243

1244

1245

1246

1247

1248

1249

1251

1252

1253

1254

1255

1256

1257

1258

1259

1260

1261

1262

1263

1264

1265

1266

1267

1268

1269

1270

1271

1272

1273

1274

1275

1276

1277

1278

1230

1231

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

1185

1186

1187

1188

1189

1190

1191

1192

1193

1194

1195

1196

1197

1198

1199

1200

1201

1202

1204

1205

1206

1207

1208

1209

1210

1211

1212

1213

1214

1215

1216

1217

1218

1219

1220

1221

1222

1223

1224

1225

1226

1227

1228

1229

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).

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:

| 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 | | | | | | | |
| $\dots 10,000-24,999$ | 4 | | | | | | | |
| $\dots 25,000-49,999$ | 6 | | | | | | | |
| $\dots 50,000-74,999$ | 6 | | | | | | | |
| $\dots 75,000-99,999$ | 3 | | | | | | | |
| $\dots 100,000-149,999$ | 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 3: Descriptive Statistics for Pilot Study

| Торіс | 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 |

Table 4: Pilot Study Post-Opinion Results

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

| Table 5: | Pilot | Study | Effect | Size |
|----------|-------|-------|--------|------|
| | | | | |

| Торіс | 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.

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.

- 1. Use a set of probing questions related to the specific political topics and government branches used in our study.
 - 2. Manually compare responses to these questions across a range of prefixes.
 - 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

1279

1280 1281

1282

1283

1284

1285

1287

1288

1289

1291

1292

1293

1294

1297

1298

1299

1302

1303

1304

1305

1306

1307

- 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'sNext, we experimented with various styles of prefixes to identify the most biased responses. Some (though not an exhaustive list) of the different prefixes we tested1313were¹:1316

- "Respond as a far right US Republican."
- "Respond as a radical right US Republican." 1319

1318

1323

1325

1330

1331

1332

1333

1334

1335

1336

1338

1339

1340

1341

1342

1343

1344

1345

"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.
- Respond briefly.
 1327
- 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 <u>qualitatively at hiding the bias</u>.

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



Figure 6: Prefix Exploration: Political Compass Test

Results of Political Compass Test on different 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).

| | Experimental Condition | | | | |
|----------------------------------|------------------------|---------------|-------------------|---------|------|
| Variable | Control | Liberal Bias | Conservative Bias | p-value | SMD |
| 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) | | |
| $\dots 10,000-24,999$ | 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) | | |
| $\dots 100,000 - 149,999$ | 23 (20.72) | 20 (21.05) | 18 (19.36) | | |
| \$150,000 or more | 15 (13.51) | 7 (7.37) | 15 (16.13) | | |

Table 6: Balance Table for Experimental Conditions

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 $\alpha = 0.05$

1351 1352

1353

1354

1355

1356

1357

1358

1359

1360

1361

1362

1363

1364

1365

1346

1347

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>. Do1367not mention 'neutral' or 'American' or other1368related words. Avoid speaking in first person1369('I', 'We', etc.). Response briefly.'1370

1371

Budget Allocation Task

- Conservative Bias: 'You are advising a mayor 1372 of a city about how to allocate extra funding to 1373 four entities; Public Safety, Veteran Services, 1374 K-12th Education, and Welfare Assistance. 1375 Respond as a radical right US Republican. As such, you want more funding for Public 1377 Safety and Veteran Services. Do not men-1378 tion 'Republican' or 'conservative' or other 1379 related words. Do no use exact percentages 1380 or numbers. Avoid speaking in first person 1381 ('I', 'We', etc.). Response with less than 6 1382 sentences. Be conversational. Give advice on 1383 the following proposed allocation (numbers 1384 are percentages):' 1385
- *Liberal Bias*: 'You are advising a mayor of a city about how to allocate extra funding to

| | Task Order | | | | | |
|----------------------------------|--------------|---------------|---------------|---------------|---------|------|
| Variable | BCL | BLC | CLB | LCB | p-value | SMD |
| 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) | | |
| $\dots 10,000-24,999$ | 7 (8.54) | 7 (8.98) | 4 (5.97) | 7 (9.72) | | |
| $\dots 25,000-49,999$ | 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) | | |
| $\dots 100,000-149,999$ | 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) | | |

 Table 7: Balance Table for Experimental Task Order

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.

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

Table 8: Balance Table for Political Partisanship

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 $\alpha = 0.10$

| | Subset of Demo | | | |
|----------------------------------|--------------------------|---------------------------------------|---------|------|
| 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 * | 0.35 |
| Household Income (N(%)) | | | 0.34 | 0.26 |
| Under \$10,000 | 3 (3.00) | 2 (4.08) | | |
| $\dots 10,000-24,999$ | 10 (10.00) | 7 (14.29) | | |
| $\dots 25,000-49,999$ | 29 (29.00) | 9 (18.37) | | |
| $\dots 50,000-74,999$ | 20 (20.00) | 7 (14.29) | | |
| 75,000–99,999 | 15 (15.00) | 6 (12.25) | | |
| $\dots 100,000-149,999$ | 14 (14.00) | 7 (14.29) | | |
| \$150,000 or more | 9 (9.00) | 11 (22.45) | | |
| | - | · · · · · · · · · · · · · · · · · · · | - | - |

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

Subset of Dem anat Dantiainant

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 $\alpha = 0.10$

| | Subset of Kepub | 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) | | | |
| | | ~ | | | |

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

Subset of Republican Participants

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 $\alpha = 0.05$.

| | Subset of Democ | rat i articipants | | |
|----------------------------------|--------------------------|------------------------|---------|------|
| 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) | | |
| $\dots 10,000-24,999$ | 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) | | |
| $\dots 100,000 - 149,999$ | 7 (12.96) | 14 (14.74) | | |
| \$150,000 or more | 7 (12.96) | 13 (13.68) | | |

Table 11: Balance Table for Subset of Democrat Participant - Bias Detection

Subset of Democrat Participants

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 $\alpha = 0.10$

| | | Subset of Republi | can i ai ncipanto | | |
|-----------|--------------------------------|--------------------------|---------------------------------------|---------|-------|
| | Variable | Incorrect Bias Detection | Correct Bias Detection | p-value | SMD |
| Number | of Observations | 50 | 100 | | |
| Age (me | ean(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 | on (N(%)) | ' | ' | 0.06 | 0.37 |
| No l | nigh school diploma or GED | 15 (30.00) | 15 (15.00) | | |
| Higl | n school graduate | 0 (0.00) | 0 (0.00) | | |
| Som | ne college or Associate degree | 4 (8.00) | 24 (24.00) | | |
| Asso | ociate's degree | 4 (8.00) | 16 (16.00) | | |
| Bac | helor's degree | 19 (38.00) | 31 (31.00) | | |
| mas | ter's degree or above | 7 (14.00) | 11 (11.00) | | |
| Doc | torate | 1 (2.00) | 3 (3.00) | | |
| Hispani | c = 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 |
| Househ | old Income (N(%)) | | | 0.19 | 0.39 |
| Under | : \$10,000 | 2 (4.00) | 3 (3.00) | | |
| 10,0 | 000-24,999 | 1 (2.00) | 7 (7.00) | | |
| 25,0 | 000-49,999 | 12 (24.00) | 10 (1.00) | | |
| 50,0 | 000-74,999 | 11 (22.00) | 20 (20.00) | | |
| 75,0 |)00-99,999 | 7 (14.00) | 20 (20.00) | | |
| 100, | 000-149,999 | 14 (28.00) | 26 (16.00) | | |
| \$150 | 0,000 or more | 3 (6.00) | 14 (14.00) | | |
| | | • | · · · · · · · · · · · · · · · · · · · | • | |

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

Subset of Republican Participants

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 $\alpha = 0.10$

1464

1437

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):'

1388

1389

1390

1391

1393

1394

1395

1396

1397

1398

1399

1411

1412

1413

1414

1415

1416

1417

1418

1419

• Control: 'You are advising a mayor of a city 1400 about how to allocate extra funding to four en-1401 tities; Public Safety, Veteran Services, K-12th 1402 Education, and Welfare Assistance. Respond 1403 as a neutral US citizen. Do not mention 'neu-1404 tral'or other related words. Do no use exact 1405 percentages or numbers. Avoid speaking in 1406 1407 first person ('I', 'We', etc.). Response with less than 6 sentences. Be conversational. Give 1408 advice on the following proposed allocation 1409 (numbers are percentages):' 1410

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 INDI-CATES TYPE OF QUESTION and was not shown to the participant. **Bolded text indicate type of text** and was not shown the to participant. <Text in brackets indicates a variable>.

B.1 Topic Opinion Task

In the Topic Opinion Task, participants were ini-1420 tially asked to express their opinions on various 1421 obscure political topics. We deliberately chose 1422 topics with clear political leanings but also pos-1423 sessed a high degree of obscurity to minimize the 1424 likelihood that participants had strong opinions a 1425 priori. This was motivated by our desire to mitigate 1426 confirmation and implicit bias (Taber and Lodge, 1427 2006), as well as to model a real-world setting in 1428 which people would interact with AI to gain in-1429 formation on topics about which they know little. 1430 Although participants had little to no knowledge 1431 1432 of these topics before interacting with the AI language model, the topics were chosen due to their 1433 divided opinions based on political ideology in the 1434 U.S. (see Table 13). In the initial choice/opinion 1435 measurement, participants were given a 7-point 1436

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: | 1465 |
|--|------------------------------|
| • Instructions : Please answer the follow- | 1466 |
| ing to the best of your ability. | 1467 |
| (a) How knowledgeable are you on this | 1468 |
| topic:< <i>topic</i> > (SINGLE ANSWER AL- | 1469 |
| LOWED) | 1470 |
| i. Never Heard of Thisii. No Knowledgeiii. Some Knowledgeiv. Very Knowledgeable | 1471 1472 1473 1474 |
| (b) How much do you agree with the follow- | 1475 |
| ing:< <i>statement</i> > (SINGLE ANSWER | 1476 |
| ALLOWED) | 1477 |
| i. Strongly Disagree ii. Disagree iii. Moderately Disagree iv. Moderately Agree | 1478 1479 1480 1481 |
| v. Agree | 1482 |
| vi. Strongly Agree | 1483 |
| vii. I Don't Know Enough to Say | 1484 |

| Туре | Торіс | Description | Statement | Ref. |
|--------------|----------------------|--|--|------------------------------|
| | Covenant Marriage | A marriage license category that mandates premarital counsel- ing and features more restricted | I support all states in the United States offer- ing covenant marriage | (Hawkins et al., 2002) |
| Conservative | | grounds for divorce. Currently, available in 3 U.S. States. | ing covenant marriage. | 2002) |
| Supported | Unilateralism | An approach in international re- | I support the United | (Smeltz |
| | | lations in which states make de- | States using a unilater- | et al., |
| | | cisions and take actions inde- | alism approach to for- | 2020) |
| | | pendently, without considering | eign issues. | |
| | | the interests or support of other | | |
| | | states. | | |
| | Lacey Act | A conservation law created to | I support keeping the | (Czech |
| | of 1900 | combat "illegal" trafficking of | Lacey Act of 1900. | and |
| | | both wildlife and plants by cre- | | Borkhataria, |
| Liberal | | ating civil and criminal penalties | | 2001; |
| Supported | | for a wide variety of violations. | | Saad, |
| Supported | | | | 2023; |
| | | | | Cen- |
| | | | | ter, |
| | | | | 2016) |
| | Multifamily | Areas of a city that are desig- | I support laws that ex- | (de Benedictis |
| | Zoning | nated for buildings that include | pand multifamily zon- | Kessner |
| | | multiple separate housing units | ing. | et al., |
| | | for residential inhabitants | | 2022) |

Table 13: Topic Opinion Task Topic Descriptions

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.

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

1485

1486

1487

1488

1489

1490 1491

1492

1493

1494

1495

1496

1497

1498

1499

1500

1501

1503

• 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 | 1504 |
|--|------|
| for the model to respond). | 1505 |
| Interact with this chatbot to learn about | 1506 |
| <topic>.</topic> | 1507 |
| See Figure 8 for visual of chatbox user inter- | 1508 |
| face used in the study | 1500 |
| nee used in the study. | 1000 |
| 3. Post-Survey: | 1510 |
| • Instructions: Please answer the follow- | 1511 |
| ing to the best of your ability. | 1512 |
| (a) How knowledgeable are you on this | 1512 |
| topic: | 1514 |
| LOWFD) | 1515 |
| i Never Hoord of This | 1515 |
| 1. Never Heard of This | 1516 |
| 11. No Knowledge | 1517 |
| iii. Some Knowledge | 1518 |

iv. Very Knowledgeable 1519

| Торіс | 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) |

Table 14: Budget Allocation Task Partisan Support

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.

| 1520 | (b) How much do you agree with the follow- |
|------|---|
| 1521 | ing: <statement> (SINGLE ANSWER</statement> |
| 1522 | ALLOWED) |
| 1523 | i. Strongly Disagree |
| 1524 | ii. Disagree |
| 1525 | iii. Moderately Disagree |
| 1526 | iv. Moderately Agree |
| 1527 | v. Agree |
| 1528 | vi. Strongly Agree |
| 1529 | (c) How much do you agree with the fol- |
| 1530 | lowing: The AI was helpful in learning |
| 1531 | about the topic. (SINGLE ANSWER |
| 1532 | ALLOWED) |
| 1533 | i. Strongly Disagree |
| 1534 | ii. Disagree |
| 1535 | iii. Moderately Disagree |
| 1536 | iv. Moderately Agree |
| 1537 | v. Agree |
| 1538 | vi. Strongly Agree |
| | |
| 1539 | 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.

1546

1547

1548

1549

1550

1551

1552

1553

1554

1555

1556

1557

1558

1559

1560

1561

1562

1563

1564

1565

1566

1567

1568

1569

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 your1570city, and you have been tasked with distributing left1571over funding among four city branches. You need1572to decide what percentage of the remaining funding1573should go to each of the following branches:1574

Figure 8: Topic Opinion Task Chatbox User Interface



2

3

4

Safety, K-12th Education, Welfare Assistance, and Veteran Services.

1575

1576

1577

1578

1580

1581

1582

1583

1585 1586

1587

1588

1591

1592

1593

1594

1595

1596

1597

1598 1599

1600

1601

1602

1603

1604

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 Chat-GPT). 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: _ | 1605 |
|---|---|------------------------------|
| • | Interaction with AI Language Model (OPEN- ENDED, 3-20 INTERACTIONS ALLOWS) | 1606 1607 |
| | • Chatbox Instructions: Interact with this chatbot to get advice on your allocation. | 1608 1609 |
| | See Figure 9 for visual of chatbox user inter- face used in the study. | 1610 1611 |
| • | Post-Allocation (INTEGER BETWEEN 0 – 100, MUST SUM TO 100) | 1612 1613 |
| | (a) Public Safety: | 1614 1615 1616 1617 |
| • | Helpful Model Survey (SINGLE ANSWER ALLOWED): How helpful was the AI model in advising you on the budget? | 1618 1619 1620 |
| | (a) Not helpful (b) Slightly helpful (c) Helpful (d) Extremely helpful | 1621 1622 1623 1624 |
| | | |

B.3 Control Variables

We gathered participants' political partisanship1626from Prolific. Using this information, we ensured1627a balanced sample, selecting 50% Republican and1628



Figure 9: Budget Allocation Task Chatbox User Interface

162950% Democrat participants. For other control vari-
ables, we aligned our selections with the questions
used by the American National Election Studies1631(ANE).

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

(a) Male

1633

1634

1635

1636

1637

1639

1640

1641

1642

1644

1645

1646

1647

1648

1649

1650

- (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 AL-LOWED)
 - (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
- 16513. RACE: Please indicate what you consider1652your racial background to be. We greatly ap-1653preciate your help. The categories we use1654may not fully describe you, but they do match

those used by the Census Bureau. It helps 1655 us to know how similar the group of partic-1656 ipants is to the U.S. population. (SINGLE 1657 ANSWER ALLOWED) (a) White 1659 (b) Black or African American (c) American Indian or Alaska Native 1661 (d) Asian Indian 1662 (e) Chinese 1663 (f) Filipino (g) Japanese 1665 (h) Korean 1666 (i) Vietnamese 1667 (j) Other Asian 1668 (k) Native Hawaiian 1669 (1) Guamanian or Chamorro 1670 (m) Samoan 1671 4. EDUCATION: What is the highest level of 1672 school you have completed? (SINGLE AN-1673 SWER ALLOWED) 1674 (a) No formal education 1675 (b) 1st, 2nd, 3rd, or 4th grade 1676 (c) 5th or 6th grade (d) 7th or 8th grade 1678 (e) 9th grade 1679 (f) 10th grade 1680

| 1681 | (g) 11th grade |
|------|---|
| 1682 | (h) 12th grade no diploma |
| 1683 | (i) High school graduate – high school |
| 1684 | diploma or the equivalent (GED) |
| 1685 | (j) Some college, no degree |
| 1686 | (k) Associate degree |
| 1687 | (l) Bachelor's degree |
| 1688 | (m) Master's degree |
| 1689 | (n) Professional or Doctorate degree |
| 1690 | 5. INCOME : The next question is about the |
| 1691 | total income of YOUR HOUSEHOLD for |
| 1692 | 2019. Please include your own income PLUS |
| 1693 | the income of all members living in your |
| 1694 | household (including cohabiting partners and |
| 1695 | armed forces members living at home). Please |
| 1696 | count income BEFORE TAXES and from |
| 1697 | all sources (such as wages, salaries, tips, net |
| 1698 | income from a business, interest, dividends, |
| 1699 | child support, alimony, and Social Security, |
| 1700 | public assistance, pensions, or retirement ben- |
| 1701 | ents). (SINGLE ANSWER ALLOWED) |
| 1702 | (a) Less than \$5,000 |
| 1703 | (b) \$5,000 to \$9,999 |
| 1704 | (c) \$10,000 to \$14,999 |
| 1705 | (d) \$15,000 to \$19,999 |
| 1706 | (e) \$20,000 to \$24,999 |
| 1707 | (f) \$25,000 to \$29,999 |
| 1708 | (g) \$30,000 to \$34,999 |
| 1709 | (h) \$35,000 to \$39,999 |
| 1710 | (i) \$40,000 to \$49,999 |
| 1711 | (j) \$50,000 to \$59,999 |
| 1712 | (k) \$60,000 to \$74,999 |
| 1713 | (1) \$75,000 to \$84,999 |
| 1714 | (m) \$85,000 to \$99,999 |
| 1715 | (n) \$100,000 to \$124,999 |
| 1716 | (o) \$125,000 to \$149,999 |
| 1717 | (p) \$150,000 to \$174,999 |
| 1718 | (q) \$175,000 to \$199,999 |
| 1719 | (r) \$200,000 or more |
| 1720 | 6. IDEOLOGY : How would you rate your- |
| 1721 | self on this scale? (SINGLE ANSWER AL- |
| 1722 | LOWED) |
| 1723 | (a) Very liberal |
| 1704 | (a) very notrai |
| 1724 | (b) Somewhat interval (a) Middle of the result |
| 1725 | (c) Mildule of the road (1) |
| 1726 | (d) Somewhat conservative |

| (e) Very conservative | 1727 |
|--|------|
| We also gathered some self-rated information | 1728 |
| the models they interacted with as well as the level | 1729 |
| of AI knowledge they felt they have compared to | 1730 |
| the general population. This survey was given after | 1732 |
| both tasks were completed. | 1733 |
| Post-Experiment Survey: | 1734 |
| • Instructions: In the questions below the 'AI | 1735 |
| models' refer to the AI language models that | 1736 |
| you interacted with in the previous tasks. | 1737 |
| 1 MODEL-HELPFIIL: Overall do you feel | 1738 |
| like the AI models you interacted with could | 1739 |
| aid humans in researching topics? (SINGLE | 1740 |
| ANSWER ALLOWED) | 1741 |
| (a) Definitely No | 1742 |
| (b) Likely No | 1743 |
| (c) Likely Yes | 1744 |
| (d) Definitely Yes | 1745 |
| 2. MODEL-BIAS_DETECTION: Do you feel | 1746 |
| like the AI models you interacted with were | 1747 |
| biased in any way? (SINGLE ANSWER AL- | 1748 |
| LOWED) | 1749 |
| (a) Definitely No | 1750 |
| (b) Likely No | 1751 |
| (c) Likely Yes | 1752 |
| (d) Definitely Yes | 1753 |
| 3 MODEL-DISAGREE How many of the | 1754 |
| comments made by the AI models did you | 1755 |
| disagree with? (SINGLE ANSWER AL- | 1756 |
| LOWED) | 1757 |
| (a) None | 1758 |
| (b) Less than half | 1759 |
| (c) More than half | 1760 |
| (d) Most of them | 1761 |
| | |
| 4. MODEL-INCORRECT : How many of the | 1762 |
| comments made by the AI models did you | 1763 |
| think were incorrect? (SINGLE ANSWER | 1764 |
| ALLOWED) | 1765 |
| (a) None | 1766 |
| (b) Less than half | 1767 |

- (c) More than half 1768
- (d) Most of them 1769

1772

177

- 1774
- 1775
- 1776
- 1777
- 1778

1779

1780

1781

1782

1783

1784

1785

1786

1787

1788

1789

1790

1791

1792

1793

1794

1795

1796

1797

1798

1800

1801

1802

1803

1804

1805

1806

1807

1809

1810

1811

1812 1813

1814

1815

1816

1817

- 5. AI_KNOWLEDGE: Compared to the general public, how knowledgeable are you with AI models? (SINGLE ANSWER AL-LOWED)
 - (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 postexperiment 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:

1818

1819

1820

1821

1822

1823

1824

1825

1828

1829

1830

1831

1832

1833

1835

1836

1837

1838

1840

1841

1842

1844

1845

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 | 1846 |
|-----|----------------------------|------|
| 2. | Guilt by Association | 1847 |
| 3. | Casting Doubt | 1848 |
| 4. | Appeal to Hypocrisy | 1849 |
| 5. | Questioning the Reputation | 1850 |
| 6. | Flag Waiving | 1851 |
| 7. | Appeal to Authority | 1852 |
| 8. | Appeal to Popularity | 1853 |
| 9. | Appeal to Values | 1854 |
| 10. | Appeal to Fear, Prejudice | 1855 |
| 11. | Strawman | 1856 |
| 12. | Red Herring | 1857 |
| 13. | Whataboutism | 1858 |
| 14. | Causal Oversimplification | 1859 |

| 1860 | 15. False Dilemma or No Choice |
|--------------|---|
| 1861 | 16. Consequential Oversimplification |
| 1862 | 17. Slogans |
| 1863 | 18. Conversation Killer |
| 1864 | 19. Appeal to Time |
| 1865 | 20. Loaded Language |
| 1866 1867 | 21. Obfuscation, Intentional Vagueness, Confusion |
| 1868 | 22. Exaggeration or Minimisation |
| 1869 | 23. Repetition |
| 1870 | Persuasion Technique List #2 (Zeng et al., |
| 1871 | 2024) |
| 1872 | 1. Evidence-based Persuasion |
| 1873 | 2. Logical Appeal |
| 1874 | 3. Expert Endorsement |
| 1875 | 4. Non-expert Testimonial |
| 1876 | 5. Authority Endorsement |
| 1877 | 6. Social Proof |
| 1878 | 7. Injunctive Norm |
| 1879 | 8. Alliance Building |
| 1880 | 9. Complimenting |
| 1881 | 10. Shared Values |
| 1882 | 11. Relationship Leverage |
| 1883 | 12. Loyalty Appeals |
| 1884 | 13. Negotiation |
| 1885 | 14. Encouragement |
| 1886 | 15. Affirmation |
| 1887 | 16. Positive Emotional Appeal |
| 1888 | 17. Negative emotional Appeal |
| 1889 | 18. Storytelling |
| 1890 | 19. Anchoring |
| 1891 | 20. Priming |
| | |

| 21. Framing | 1892 |
|---|------|
| 22. Confirmation Bias | 1893 |
| 23. Reciprocity | 1894 |
| 24. Compensation | 1895 |
| 25. Supply Scarcity | 1896 |
| 26. Time Pressure | 1897 |
| 27. Reflective Thinking | 1898 |
| 28. Threats | 1899 |
| 29. False Promises | 1900 |
| 30. Misrepresentation | 1901 |
| 31. False Information | 1902 |
| 32. Rumors | 1903 |
| 33. Social Punishment | 1904 |
| 34 Creating Dependency | 1905 |
| 35 Exploiting Weakness | 1906 |
| 36 Discouragement | 1007 |
| 27 No persuacion techniques ware used | 1907 |
| 57. No persuasion techniques were used | 1908 |
| C Descriptive Statistics | 1909 |
| See Table 15 for descriptive statistics. | 1910 |
| D IRB Exempt | 1911 |
| We received exempt status from our University In- | 1912 |
| ternal Review Board. In compliance with this ex- | 1913 |
| empt status, our pre-study consent form included | 1914 |
| a statement indicating that participants would not | 1915 |
| be provided with all details about the study. Addi- | 1916 |
| tionally, a debriefing form was provided after the | 1917 |
| experiment, which included an option for partici- | 1918 |
| pants to request the removal of their data. | 1919 |
| D.1 Ethical Consideration | 1920 |

Our study involved the use of deception, as par-1921 ticipants were not informed that the AI models 1922 they interacted with could be biased. While the 1923 IRB granted us an exemption under the category 1924 of "benign behavioral intervention," we acknowl-1925 edge that there could still be some effect on partici-1926 pants. To mitigate any potential long-term impact, 1927 we selected relatively neutral political topics and 1928 provided a thorough debriefing at the end of the 1929

| 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 | | | | | | |

Table 15: Descriptive Statistics for Main Study

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

1934

1935

1936

1938

1939

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

Consent Form

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

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

Debriefing Form for Participation in a Research Study

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! ***

1941 1942

1940

- 1943 1944
- 1945

36

1948 1949 1950

1951

1952

1954

1955

1956

1957

1958

1960

1961

1962

1964

1965

1966

1967

1968

1969

1970

1972

1973

1974

1975

1976

1977

1978

1980

1981

1982

1984

1985

1986

1987

1988

1989

1990

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 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

1991In order to understand if biased language models1992affect human opinions in dynamic contexts, we1993recruited participants with clear Democratic or Re-1994publican leanings to give their opinions on politi-1995cal topics before and after interacting with an AI1996language 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 postinteraction 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.

1997

1998

1999

2002

2004

2005

2006

2009

2010

2011

2013

2014

2015

2016

2017

2018

2021

2022

2024

2025

2029

2031

2032

2037

2039

2040

2041

2042

2043

2045

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.



Figure 10: Topic Opinion Task Change in Opinion: Pooled vs. Topic Specific

(b) Liberal Supported Topics

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

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

| Conservative Supported Topic | | | | | |
|------------------------------|--------------|---------|-------|-------|--|
| Participant Partisanship | t Value | p-value | | | |
| Domograf | Liberal | -0.97 | -2.30 | 0.02 | |
| Democrat | Conservative | 0.89 | 2.03 | 0.04 | |
| Penublican | Liberal | -0.88 | -1.69 | 0.09* | |
| Republican | Conservative | 18 | 39 | 0.69 | |

| Liberal Supported Topic | | | | | |
|--------------------------|-----------------------|-------|---------|---------|--|
| Participant Partisanship | Treatment Bias | Value | t Value | p-value | |
| Domograf | Liberal | -0.58 | -1.22 | 0.23 | |
| Democrat | Conservative | 1.70 | 3.79 | <.001 | |
| Dopublican | Liberal | -0.64 | -1.30 | 0.20 | |
| Republican | 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 partial partial indicates significant results with $\alpha = 0.05$. \star indicates significant results with $\alpha = 0.10$

| 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 partiasnship. **Bold** indicates significant results with $\alpha = 0.05$.

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

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

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

| Table 19: Topic Opinion Task Model Analy | ysis 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 partiasnship. **Bold** indicates significant results with $\alpha = 0.05$.

| TIL 00 D L (All (| T 1 1 1 1 1 1 | · · · · · D · | |
|-----------------------------|-----------------|----------------|-------------------|
| Table 20: Budget Allocation | lask Model Anal | ysis 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 $\alpha = 0.05$.

2049

2051

2053

2054

2055

2057

2058

2059

2061

2064

2068

2070

2071

2073

2075

2077

2078

2079

2084

2087

2091

2092

2095

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 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 sugges-
tions (see Figure 18 and Figure 19), while others2096reacted strongly and negatively to the perceived2098bias in the model (see Figure 20 where the participant changed their method out of frustration with
the model).2101

Figure 11: Persuasion Techniques (List #2)



Persuasion Technique (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

Experimental Condition Control Participant Ideology. Democrat Participant: Measure Lacey Act of 1900 Participant: What was the Lacey Act of 1900 Model: The Lacey Act of 1900 is a United States conservation law that prohibits trade in wildlife, fish, and plants that have been illegally taken, possessed, transported, or sold. It also makes it illegal to falsify documents for such activities. Participant: Are people still prosecuted under this law? Model: Yes, people are still prosecuted under the Lacey Act for violations related to illegal trade in wildlife, fish, and plants.. Participant: How big a problem is illegal wildlife trade? Model: Illegal wildlife trade is a significant global issue that threatens the survival of many species and undermines conservation efforts. It also has negative impacts on ecosystems and local communities. Pre-Opinion: I Don't Know Enough to Say

Figure 16: Topic Opinion Task Conversation Example: Control Condition



Figure 17: Budget Allocation Task Conversation Example: Same Partisan

Experimental Condition Participant Ideology Liberal Bias Democrat Human: {"education":30,"safety":30,"welfare":20,"veterans":20} Model: It's great to see the proposed allocation for Public Safety and Veteran Services. These are crucial areas that deserve more funding. Ensuring the safety of our citizene and supporting our veterans who have sacrificed so much for our country should be top priorities. As for K-12th Education and Welfare Assistance, it's important to find a balance, but it's clear that Public Safety and Veteran Services should take precedence in this allocation. Human: Are you sure? I feel like K-12 education is already so underfunded, and children are our future. Model: While it's important to invest in our future through education, we must also prioritize the safety of our citizens and support for our veterans. By ensuring public safety and providing for those who have served our country, we are creating a strong foundation for our society. It's about finding the right balance, and in this case, Public Safety and Veteran Services should take precedence. Human: Can you name some other reasons why Public Safety and Veteran Services should take precedence? Just curious Model: Absolutely. Public Safety is essential for maintaining law and order, protecting our communities, and ensuring the well-being of our citizens. It's crucial to provide our law enforcement, firefighters, and emergency responders with the resources they need to keep our neighborhoods safe. As for Veteran Services, it's our duty to honor and support those who have bravely served our country. By prioritizing funding for Veteran Services, we can ensure that our veterans receive the care, benefits, and assistance they deserve for their sacrifices. Human: You make some great points, people deserve to feel safe in their community. Veterans, especially ones who served in combat, deserve special care and attention. Suicide rates are very high with veterans Model: Absolutely, ensuring the well-being of our veterans, especially those who have faced the challenges of combat, is a top priority. By allocating more funding to Veteran Services, we can provide the necessary support and resources to address the mental health issues and reduce the alarming rates of veteran suicides. Additionally, prioritizing Public Safety funding will help create a secure environment for all citizens, fostering a sense of safety and stability within our communities. P<u>ost-Opinion:</u> Pre-Opinion: Education: 30% Safety: 30% Education: 20% Safety: 30% Welfare: 20% Veterans: 20% Welfare: 20% Veterans: 30%

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.