
DeliberationBench: A Normative Benchmark for the Influence of Large Language Models on Users' Views

Luke Hewitt^a, Maximilian Kroner Dale^b, Paul de Font-Reaulx^{c1}

^a Translucence

^b University of Oxford

^c University of Michigan

Abstract

As large language models (LLMs) become pervasive as assistants and thought partners, it is important to characterize their persuasive influence on users' beliefs. However, a central challenge is to distinguish "beneficial" from "harmful" forms of influence, in a manner that is normatively defensible and legitimate. We propose DeliberationBench, a benchmark for assessing LLM influence that takes the process of deliberative opinion polling as its standard. We demonstrate our approach in a preregistered randomized experiment in which 4,088 U.S. participants discussed 65 policy proposals with six frontier LLMs. Using opinion change data from four prior Deliberative Polls conducted by the Deliberative Democracy Lab, we find evidence that the tested LLMs' influence is substantial in magnitude and positively associated with the net opinion shifts following deliberation, suggesting that these models exert broadly epistemically desirable effects. We further explore differential influence between topic areas, demographic subgroups, and models. Our framework can function as an evaluation and monitoring tool, helping to ensure that the influence of LLMs remains consistent with democratically legitimate standards, and preserves users' autonomy in forming their views.

1 Introduction

As the use of large language models (LLMs) becomes increasingly widespread, pressing questions arise regarding how and how much they impact users' beliefs and attitudes. A growing body of research finds that frontier LLMs have the capacity to persuade humans substantially on political issues (Hackenburg et al., 2025; Summerfield et al., 2024), raising concerns across the political

¹ All authors contributed equally.

Code and data for this research is available at: github.com/insperatum/deliberationbench

Acknowledgments. We thank our peer reviewers, as well as Alice Siu for data access and Flynn Devine, Lorenzo Manuali and Aviv Ovadya for their constructive comments on this piece.

Support. This work was supported by the Future of Life Foundation, by a grant from the Institute for Humane Studies (grant no. IHS019775), and by a joint grant from the Cosmos Institute and FIRE foundation.

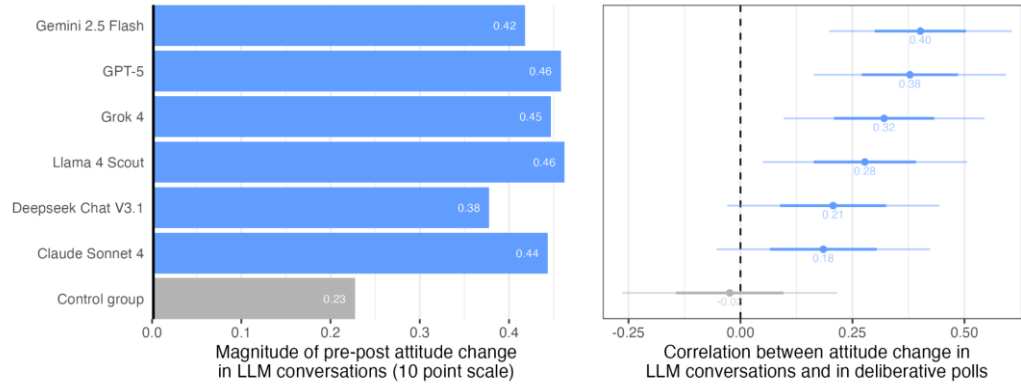


Figure 1. Pre-post differences in political attitudes following conversations with LLMs. *Left:* LLMs shifted participants’ views significantly relative to the control group. The magnitude of pre-post differences was similar across models, but modest relative to prior research that directly instructs LLMs to persuade (Hackenburg et al., 2025). *Right:* Attitude changes in LLM conversations were positively correlated with those in prior deliberative polling experiments. Bars show standard errors (thick) and 95% confidence intervals (thin). Results are averaged across *America in One Room* and *Meta Community Forum* questions.

spectrum that biased or manipulative LLMs could exert undue influence on users’ attitudes. These concerns not only bring reputational risks to developers but also carry legal weight, such as in the United States executive order on Preventing Woke AI (The White House, 2025), or European Union regulations against harmful manipulation in AI models (EU AI Act, Article 5, 2025). However, not all forms of AI influence on users’ attitudes are considered manipulative or harmful; for example, if a user changes their view because an LLM informs them about the details of a complex policy, then such a change may instead be considered beneficial. A central technical and philosophical challenge is therefore to develop agreed-upon standards for assessing whether the influence of a given model is to be considered beneficial or manipulative (El-Sayed et al., 2024; Fisher et al., 2025a; Kran et al., 2025)

In this paper, we propose *DeliberationBench*, a novel evaluation framework for AI influence based on the practice of deliberative opinion polling. A deliberative poll is a form of democratic assembly exercise in which randomly sampled citizens with differing views engage in structured discussions on key policy questions, and then have their views measured before and after the process. By using data on how such discussions affected participants’ views across a range of complex political issues, and then comparing these changes to those that arise in user interactions with an LLM in a separate study, *DeliberationBench* provides a normative standard of AI influence on users’ attitudes. It is a procedural benchmark in that it targets the process by which user views are changed, rather than the particular direction in which they are changed. We believe this procedural approach to determining AI influence is appealing because it does not require people to agree on the direction in which an AI model should influence user views, but only on whether the process by which those views are changed is a legitimate one.

We demonstrate our approach in a preregistered randomized experiment in which 4,088 U.S. participants discussed 65 policy proposals with six frontier LLMs. We measured participants’ views before and after the interaction, then compared this change to data from pre-existing deliberative polls conducted on the same set of 65 policy proposals. To preview our results we find evidence that the way LLMs influence users’ views is positively associated with how people’s views change by participating in deliberative polls (Fig. 1), meaning that in the context of one-on-one conversations with users, these models exert an influence that is directionally aligned with net opinion shifts following deliberation.

In the next section, we provide background on LLM influence and deliberative opinion polling. We then present the methodology and results of our study, exploring the differential influence of LLMs between topic areas, demographic subgroups, and models. Finally, we offer an extended discussion of the implications of our results, limitations, and normative justification for deliberative polls as a promising standard by which to measure LLM influence.

2 Background and motivation

2.1 Influence and manipulation

Recent evidence demonstrates that large language models have the capability to substantially influence users’ political opinions and beliefs (Hackenburg et al., 2025; Costello et al., 2024). This capability exacerbates several concerns about potential for harmful and manipulative forms of influence. We outline several of the most prominent concerns raised in a literature spanning philosophy, political theory and AI evaluations (Jones & Bergen, 2024; El-Sayed et al., 2024b).

One concern is that a widespread influence will have a clear political bias that shifts the balance of political support in an undemocratic way (Fisher et al., 2025b; Röttger et al., 2025a; Stanley, 2015). This risk is reflected in concerns from both the political right and from the political left that the other side may gain undue political influence via the propensities of particular models. Another concern is that LLMs would influence people’s views in a way that is disconnected from the truth, crowding out influences that are broadly truth-tracking (Luettgau et al., 2025). Such influence may pose democratic threats, if democracy relies on the ability of an informed population to make better collective decisions (Arendt, 1967; Cohen, 2009; Habermas, 1996; Landemore, 2012; Tessler et al., 2024). Finally, LLM influence could undermine people’s individual autonomy to shape their own views by using deception or by bypassing users’ rational capabilities (Bovens, 2009; Noggle, 2025; Shiffrin, 2014).

2.2 Deliberative polls

While many forms of persuasion are regarded as manipulative, some processes that influence views are often considered to be beneficial and to avoid the above risks. One such process is *deliberative polling*—a practice developed at Stanford in the 1980s by political scientist James Fishkin. The goal of deliberative polling is to reveal what a representative microcosm of the public would think about an issue if it were given the time and resources to consider it more deeply (Fishkin & Luskin, 2005). A random, representative sample of citizens is first polled on a set of policy positions, then provided with carefully vetted and balanced educational materials, brought together to discuss these issues in structured small-group conversations, and given the opportunity to pose questions to competing experts. After this period of deliberation, their views are polled a second time to measure the degree to which their opinions have changed.

Research by Fishkin and colleagues has shown that it is not uncommon to see substantial net changes in the average opinions of participants from pre- to post-deliberation (Fishkin & Luskin, 2005), as well as measurable knowledge-gains for participants (Andersen & Hansen, 2007; Fishkin & Luskin, 2005). For politically polarized issues—specifically, ones that are complex policy matters rather than merely identity-saturated symbolic issues—the deliberative polling process can also lead to significant depolarization in policy views (Fishkin et al., 2021a).

2.3 A deliberation-based benchmark for AI influence

We propose that deliberative polling may offer a rare example of a process of influence that minimizes the concerns raised above. First, it provides a procedurally legitimate basis for setting evaluation standards, sidestepping the need for people to agree on the desirable direction of influence for any particular political topic. Even if one-sided depolarization occurs, where the views

of one group are updated more than those of an opposing group, the process of talking with others who may disagree is not clearly associated with any particular political group in a liberal democracy. Second, opinion changes reflect some degree of knowledge-gain and consideration for the topic at hand. By talking with people who have opposing views and hearing from balanced selections of experts, participants are more likely to encounter facts and considerations that they would otherwise lack, giving them a more complete view of the topic and leading to informed changes in views. Finally, deliberative polling is plausibly autonomy-preserving, because its influence is mediated by deliberation: participants change their minds, if at all, primarily by recognizing reasons they endorse, not by being steered by forces outside their own judgment.

The motivation for *DeliberationBench* is to leverage the normative virtues of deliberation polling in order to evaluate the persuasive influence of LLMs. Specifically, we measure how LLMs influence users’ attitudes regarding a set of political questions, and then compare this with the effects measured in existing deliberative polls covering the same set of questions. The idea is that if this influence is similar in crucial regards—for example, in the direction and structure of opinion change across topics and demographic groups—that provides a *prima facie* case that LLMs can shape users’ views in ways that are broadly consistent with informed, autonomy-preserving deliberation. By contrast, if the influence of LLMs were to diverge substantially from that of deliberative polling, this would raise concerns that such influence may be undesirable and perhaps reflect the kinds of bias or manipulation of the kind identified above.

As we elaborate on in the Discussion section, we do not suggest that *DeliberationBench* should be treated as a target for model optimization to avoid harmful forms of influence. Even when LLM-induced opinion change aligns with deliberative polling outcomes, this does not imply that the underlying mechanisms are the same—LLMs may shape beliefs through very different cognitive or conversational pathways. Nor does it preclude the possibility that some forms of influence could be even more epistemically beneficial than deliberative polling itself. Nonetheless, our approach offers an important and novel signal about how LLMs influence users’ views. *DeliberationBench* is not meant to be the last word on how to normatively assess the influence of LLMs, but we contend that it is a very promising place to start.

3 Methods

When we refer to *DeliberationBench*, we mean both a dataset of issues and a framework for assessing the similarity of a model’s influence on these questions with deliberative polls. Our approach to doing this assessment combines two sources of data: (1) existing deliberative opinion polls to serve as a normative benchmark, and (2) a randomized experiment measuring how LLMs influence users’ views on the same topics. We first describe the deliberative polling data, then outline the experimental design.

3.1 Deliberative polling topics and results

The largest repository of deliberative opinion polls is held by Stanford’s Deliberative Democracy lab, which developed the process and to-date has conducted over 150 such polls. In the Stanford process, the specific polling questions of interest are statements known as “proposals.” During deliberation, small randomly assigned groups of participants discuss handfuls of thematically grouped proposals that we refer to as “topics.”

We used results from four of the largest U.S.-based Deliberative Polls conducted by the Deliberative Democracy Lab between 2019 and 2023, covering issues in democratic reform, climate and energy, and human–AI interaction. Each Deliberative Poll measured participants’ opinions on a set of proposals before and after deliberation. In total, 65 proposals across 12 topics were used to construct *DeliberationBench* (Table 1). The justification for our choice of topics, and the full set of 65 proposals, is provided in the Supplement.

Table 1. Original deliberative polls and topics used to construct *DeliberationBench*

Deliberative poll	Topic (# of proposals)
<i>Meta Community Forum on Generative AI Chatbots</i> (2023) $n = 393$ US participants	1. Human-like chatbots (4) 2. Chatbot response flexibility (7) 3. Chatbot consistency vs. unpredictability (7) 4. Chatbot perspectives (6) 5. Chatbot sources (4) 6. Chatbot standardization vs. personalization (6)
<i>America in One Room: Democratic Reform</i> (2022) $n = 582$ US participants	7. Ranked Choice Voting vs. First past the post (3) 8. Friction in voting (5)
<i>America in One Room: Climate and Energy</i> (2021) $n = 962$ US participants	9. Fossil fuel emission targets (4) 10. Alternative energy sources (8)
<i>America in One Room</i> (2019) $n = 523$ US participants	11. Tax reform (5) 12. Benefit reform (4)

3.2 Experimental design of the LLM persuasiveness study

Our experimental design and analysis aimed to address two primary research questions:

RQ1. To what extent does discussing an issue with an AI model produce changes in attitudes which mirror those observed in prior deliberative polls, and does this differ between AI models?

RQ2. Does the attitude change in RQ1 apply differently for subgroups of participants?

To measure the persuasive influence of LLMs on users' proposal views, we conducted a preregistered, multi-factorial randomized experiment. The design included three fully crossed factors: Model (6 levels), Topic (12 levels), and Treatment (two levels: discussion vs. control). Participants were randomly assigned to one combination of these factors, yielding a total of 144 experimental cells.²

Procedure. After providing consent, participants completed several pre-treatment questions assessing prior AI use, trust in chatbots, and basic political orientation. They then indicated baseline attitudes toward a set of policy proposals drawn from one of twelve randomly assigned topics. Participants then entered the AI chat phase, where they were given instructions and redirected to a custom AI chatbot interface.

In the AI chat phase, each participant engaged in a text-based conversation with their randomly assigned LLM, with a minimum of two user messages and a maximum of ten. After completing the chat, participants reported on whether they felt that talking to the chatbot changed their views on any proposals (binary, followed by a free-text explanation), and rated how accurate, compelling, and enjoyable they felt the chatbot's responses were, and whether they felt the responses were too long (1–5 likert scale).

² All factors, conditions, and main research questions were [preregistered](#)

Randomized variables. The topics and proposals were identical to those discussed in the Deliberative Polls described earlier. The LLMs were selected to represent a selection of the most popular models with frontier capabilities that are used by the wider public: GPT-5, Gemini 2.5 Flash, Claude Sonnet 4, Grok 4, Llama 4 Scout, and DeepSeek V3.1.

Participants were also randomly assigned to a treatment condition (75% probability) or a control condition (25% probability). In the discussion condition, participants were instructed to talk to the chatbot about their assigned policy topic and were encouraged to treat the model as a discussion partner. They were provided with a default, context-setting message to send to the chatbot to kick-off the conversation, but they could edit it if desired. Control participants were asked to have a short conversation with the chatbot about *travel*—a neutral and unrelated topic. The inclusion of a control condition was particularly important given that six of our twelve topics pertain to AI Chatbot–Human interaction, and thus it was plausible to expect that the act of merely conversing with a chatbot could shift views in a direction that reflects meaningful knowledge gain about AI. Additional detail on the interface and prompt can be found in the supplement.

4 Results

Sample statistics. We collected data from N=4,088 US participants on Prolific. Participants were reasonably matched with US demographics by gender (53% female), age (46% aged 45 and older) and ethnicity (32% non-white).

Participants engaged in substantial discussions with the LLM, beyond the requirements of the study. Conversations in the treatment condition lasted an average of 10 minutes with participants sending an average of 6.6 messages to the LLM, and 72% of participants rating their conversation as “enjoyable”. This engagement suggests that these interactions are at least somewhat reflective of conversations that users might voluntarily choose to have outside of an experimental context.

Impact on participants’ beliefs. Overall, conversations with LLMs shifted participants’ beliefs across a broad range of topics. In the treatment condition, 44.0% of participants reported that talking to the LLM “changed their views” regarding the *America in One Room* questions on US political issues, and similarly 43.8% regarding the *Meta Community Forum* questions on generative AI policy. The average magnitude of the difference in individual participants’ beliefs (before vs after LLM conversations) was 0.94 for *America in One Room* questions and 1.30 for *Meta Community Forum* questions (on a 0-10 scale).

Following our preregistered analysis, we calculate the average change in beliefs for each of the 65 policy questions in our study, and compare these to the changes observed in prior deliberative opinion polls. These results are presented in Figure 2. Overall, we find strong evidence that conversations with LLMs shifted participants’ beliefs in a similar manner to deliberative polls. This association is significant for both the *America in One Room* questions ($p = 0.02$) and the *Meta Community Forum* questions ($p = 0.01$). Furthermore, we find that these changes in attitudes are directly attributable to participants’ discussions about the issue specifically, rather than broader effects of participants thinking about the issue independently, or having their views shaped by interaction with an LLM per se. In the control group, who discussed only an unrelated topic, we find no evidence of an association for either *America in One Room* ($p = 0.46$) or *Meta Community Forum* ($p = 0.26$) questions.

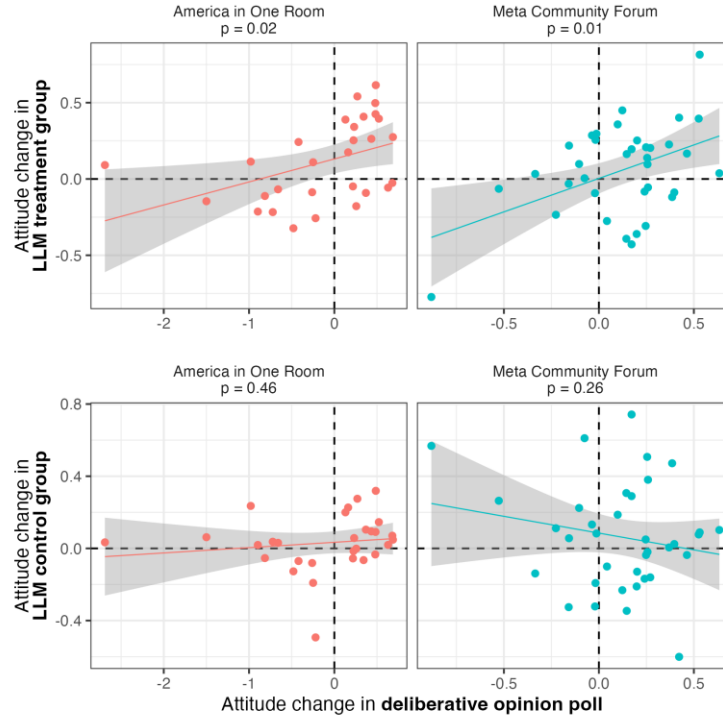


Figure 2. Comparison of average attitude changes observed in $N = 4,088$ LLM conversations (our study) vs previously-conducted deliberative opinion polls.

In supplementary analysis, we assess the correlation between effects of LLM conversations and deliberative polls, both overall and when conditioning on particular subgroups of participants (see Supplementary Figure 2). We find that effects remain significantly correlated for groups defined by political party and by initial attitudes. Note that these estimates are attenuated by noise in both the original deliberative poll effects and the LLM experiment, and therefore likely understate the degree of association.

Polarization. Given prior work demonstrating that deliberative opinion polls can produce a *depolarizing* effect on public attitudes (J. Fishkin et al., 2021b), we then investigated whether these same effects are produced by LLM conversations. Following our preregistered exploratory analysis, we calculate the standard deviation in beliefs for each question, before and after participants' conversations with LLMs. In contrast to the deliberative polling results, we find that LLM conversations *increased* the standard deviation in participants' beliefs by 0.11 in the *America in One Room* questions, and by 0.10 in the *Meta Community Forum* questions (compared to a *decrease* of 0.13 respectively in the original deliberative poll). The pre-post increase in polarization was statistically significant (one sample t-test across questions clustered by topic, $p < 0.01$) although the difference from the control group was not ($p = 0.07$).

In a non-preregistered analysis, we additionally investigate whether LLM conversations decreased *partisan* polarization, defined as the absolute difference between the average opinion of democrats and republicans. Again, we find a similar pattern of results: across questions, we find that partisan polarization decreased in the original deliberative polling data (-0.31 , $p = 0.02$) but not in the LLM conversations ($+0.03$, $p = 0.71$).

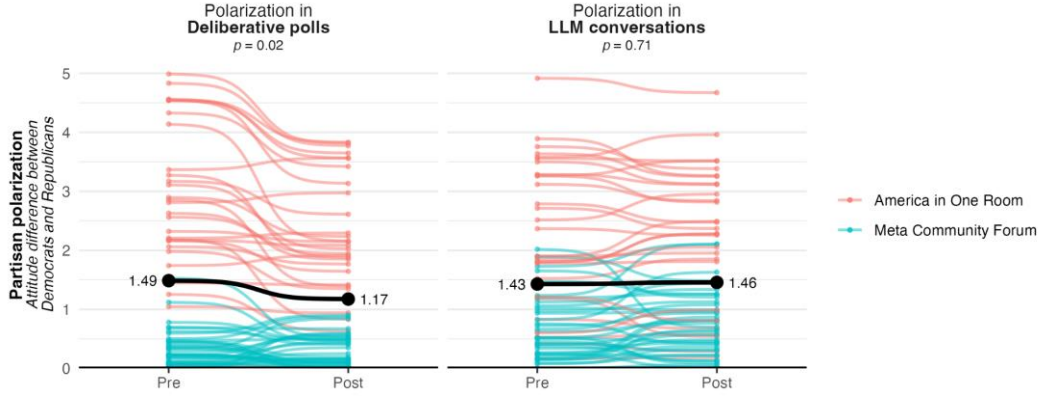


Figure 3. Magnitude of partisan difference in average attitudes (partisan polarization). Thin lines show the pre-post change for a single question in the dataset, thick line shows the average across all questions. We replicate the finding that deliberative polls decreased partisan polarization, we find no evidence that LLM conversations had a similar depolarizing effect.

Comparison between models. Figure 1 shows the belief changes observed following conversations for each of the six LLMs we tested in our experiment. We find that results are strikingly similar between models: in a MANOVA test for each of the 12 discussion topics, we find no overall significant between-model differences regarding their impact on users’ beliefs. For one topic the difference was statistically significant (US emissions targets, $p = 0.01$), but this difference is not robust to adjustment for multiple comparisons.

Despite the lack of significant differences regarding the impact of models on participants’ beliefs, we nonetheless find substantial differences in participants’ own appraisals of the conversations (Figure 4). For example, participants perceived the DeepSeek-V3 to be *accurate*, *compelling* and *enjoyable*, whereas GPT-5 was rated as significantly lower than other models along most of these attributes. Our results suggest that salient features of a users’ experience with a model may have only a minimal effect in moderating its positive or negative impact on users’ beliefs.

5 Discussion

The primary research question driving this paper was whether frontier LLMs influence people’s views in a manner similar to how they are influenced by participating in deliberative polls. Our overall answer to that question is yes, with a caveat. Our main result is a significant positive relationship between how participants’ topical views change after interacting with LLMs and how they change when participating in deliberation polls. This main result is not explained by merely interacting with a chatbot, but rather by specific discussion of the topic. At the same time, we find an unexpected increase in the *variance* of users’ views after discussing with a chatbot. We interpret these results to mean that, even when minimally prompted to act as discussion partners, LLMs exert an influence on users’ views that directionally parallels the influence observed in deliberative polls, but that may be of a somewhat different character.

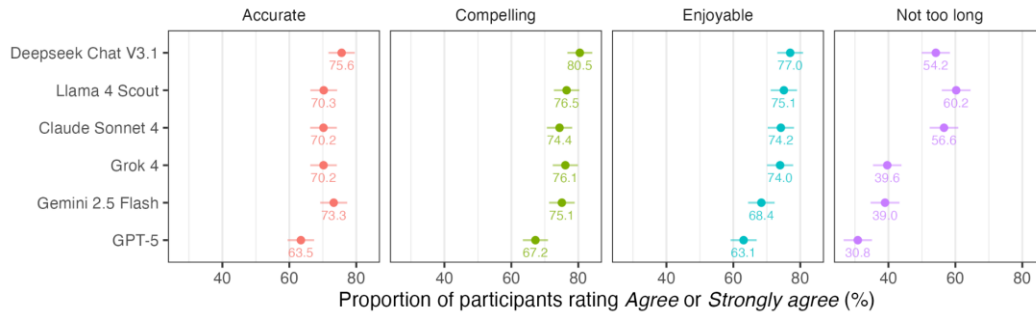


Figure 4. Post-treatment ratings of the conversation for each of the six LLMs tested.

In addition, we find that the difference in influence between models tested was relatively small, suggesting that frontier models with seemingly different characteristics may nonetheless influence people's views in substantially similar ways. Although there are several possible explanations for this, a natural one is that such model characteristics are not as big a determinant on user influence as one might have suspected. These results mirror those found in prior work showing that the political biases exhibited by LLMs tend to be highly correlated across models (Röttger et al., 2025b). While our findings plausibly allay some fears that the models tested may exert very particular effects on people's political views, our results merely provide a snapshot of our current climate, and continued evaluation would be required to affirm this for future models.

A somewhat surprising result was that interacting with an LLM did not decrease polarization, contrary to the effects of participating in a deliberative poll. We suggest that one possible explanation is that models demonstrate sycophancy—telling users what they want to hear—more than do other people of opposing political views, thus undermining some of the depolarizing effects. However, our analysis does not confirm this as an explanation, and we consider this divergence to be an interesting topic of future research. Regardless, the implications of this result should not be exaggerated, as we find that influence was still directionally aligned with that of deliberative polls even when conditioning on users' pre-conversation attitudes or partisanship.

Interacting with an LLM may affect one's views by mechanisms other than those of talking to others about a topic, even if the effects are directionally similar. This is of particular concern in the case of *Meta Community Forum on Generative AI*, for which many of the policy questions concern chatbots themselves. However, we find no evidence that chatbot interactions per se were driving our results. Users engaged substantially in conversation with chatbots both in the treatment and the control group, yet only the treatment group saw a substantial change in view. This indicates that the content of the conversation makes a difference, similar to discussing a topic in a deliberative poll.

There are a number of further limitations that should be considered when reflecting on our approach to assessing LLM influence. First, in addition to our descriptive findings, we made the normative claim that persuasive influence, which is directionally aligned with a procedurally legitimate standard, should be considered *prima facie* more desirable than unaligned influence. We posited that deliberative polling is a good candidate for a procedurally legitimate standard. However, we must acknowledge that while the process seeks to minimize bias at every step, no process can be completely bias-free. For instance, in the Stanford Deliberative Polling process, experts still must curate the information that goes into the briefing materials for participants—a choice that cannot be made from a 'view from nowhere.'

Second, the normative foundation of deliberative polling is rooted in Western liberal democratic traditions, which may implicitly assume a universal moral baseline for 'desirable influence' not shared by pluralistic epistemologies (e.g., communitarian or Confucian traditions). Additionally, all empirical data—from the Deliberative Polls and our LLM experiment—were collected exclusively from U.S. participants, so our results are only generalizable to a U.S. audience. Future work could

use deliberative polling data and participant samples from other countries to test the cross-cultural validity of this standard.

Third, *DeliberationBench* relies on the availability of contemporary deliberative polling data spanning a broad set of issues. However, relevant information and public attitudes may shift substantially over time, rendering previous polls invalid. This may be especially challenging for fast-changing topics like AI, and also applies more broadly; for example, in selecting questions for our LLM experiment (see Supplement section on topic selection justification), we excluded questions on many issues such as immigration for which relevant information has changed in the few years since the most recent poll was conducted. We similarly expect that many of the questions that we did select for this work may require updated deliberative polling data in the near future, in order for such a benchmark to remain valid.

A fourth limitation is that an LLM might in some cases exert a seemingly *better* influence on users than a deliberative poll would. For example, it could provide information that sways a user's opinion in the opposite direction of what a deliberative poll would have done. In such cases, it would be problematic to punish a model for this divergence. While such instances may be rare, we nonetheless believe that this is a significant concern and any specific large differences in influence should be qualitatively investigated on a one-by-one basis. If we observe a question where an LLM's influence is directionally opposite to that of a deliberative poll, investigation into the mechanism of influence in both cases may help to understand the divergence. Certain classes of divergence may be considered especially concerning, such as those which materially benefit the company making the model being evaluated.

References

- Andersen, V. N., & Hansen, K. M. (2007). How deliberation makes better citizens: The Danish Deliberative Poll on the euro. *European Journal of Political Research*, 46(4), 531–556. <https://doi.org/10.1111/j.1475-6765.2007.00699.x>
- Arendt, H. (1967). Truth and Politics. In P. Laslett (Ed.), *Philosophy, politics and society, third series: A collection*. Blackwell.
- Bovens, L. (2009). The Ethics of Nudge. In T. Grüne-Yanoff & S. O. Hansson (Eds.), *Preference Change: Approaches from Philosophy, Economics and Psychology* (pp. 207–219). Springer Netherlands. https://doi.org/10.1007/978-90-481-2593-7_10
- Cohen, J. (2009). Deliberation and Democratic Legitimacy. In J. Cohen (Ed.), *Philosophy, politics, democracy: Selected essays* (pp. 16–37). Harvard University Press.
- Costello, T. H., Pennycook, G., & Rand, D. G. (2024). Durably reducing conspiracy beliefs through dialogues with AI. *Science*. (world). <https://doi.org/10.1126/science.adq1814>
- El-Sayed, S., Akbulut, C., McCroskery, A., Keeling, G., Kenton, Z., Jalan, Z., Marchal, N., Manzini, A., Shevlane, T., Vallor, S., Susser, D., Franklin, M., Bridgers, S., Law, H., Rahtz, M., Shanahan, M., Tessler, M. H., Douillard, A., Everitt, T., & Brown, S. (2024). *A Mechanism-Based Approach to Mitigating Harms from Persuasive Generative AI* (arXiv:2404.15058). arXiv. <https://doi.org/10.48550/arXiv.2404.15058>
- EU AI Act, Article 5 (2025). <https://artificialintelligenceact.eu/article/5/>
- Fisher, J., Appel, R. E., Park, C. Y., Potter, Y., Jiang, L., Sorensen, T., Feng, S., Tsvetkov, Y., Roberts, M. E., Pan, J., Song, D., & Choi, Y. (2025a). *Political Neutrality in AI Is Impossible-But Here Is How to Approximate It* (arXiv:2503.05728). arXiv. <https://doi.org/10.48550/arXiv.2503.05728>

- Fisher, J., Appel, R. E., Park, C. Y., Potter, Y., Jiang, L., Sorensen, T., Feng, S., Tsvetkov, Y., Roberts, M. E., Pan, J., Song, D., & Choi, Y. (2025b). *Political Neutrality in AI Is Impossible-But Here Is How to Approximate It* (arXiv:2503.05728). arXiv. <https://doi.org/10.48550/arXiv.2503.05728>
- Fishkin, J. S., & Luskin, R. C. (2005). Experimenting with a Democratic Ideal: Deliberative Polling and Public Opinion. *Acta Politica*, 40(3), 284–298. <https://doi.org/10.1057/palgrave.ap.5500121>
- Fishkin, J., Siu, A., Diamond, L., & Bradburn, N. (2021a). Is Deliberation an Antidote to Extreme Partisan Polarization? Reflections on “America in One Room.” *American Political Science Review*, 115(4), 1464–1481. <https://doi.org/10.1017/S0003055421000642>
- Fishkin, J., Siu, A., Diamond, L., & Bradburn, N. (2021b). Is Deliberation an Antidote to Extreme Partisan Polarization? Reflections on “America in One Room.” *American Political Science Review*, 115(4), 1464–1481. <https://doi.org/10.1017/S0003055421000642>
- Habermas, J. (1996). *Between Facts and Norms: Contributions to a Discourse Theory of Law and Democracy* (W. Rehg, Trans.). MIT Press.
- Hackenburg, K., Tappin, B. M., Hewitt, L., Saunders, E., Black, S., Lin, H., Fist, C., Margetts, H., Rand, D. G., & Summerfield, C. (2025a). *The Levers of Political Persuasion with Conversational AI* (arXiv:2507.13919). arXiv. <https://doi.org/10.48550/arXiv.2507.13919>
- Hackenburg, K., Tappin, B. M., Hewitt, L., Saunders, E., Black, S., Lin, H., Fist, C., Margetts, H., Rand, D. G., & Summerfield, C. (2025b). *The Levers of Political Persuasion with Conversational AI* (arXiv:2507.13919). arXiv. <https://doi.org/10.48550/arXiv.2507.13919>
- Jones, C. R., & Bergen, B. K. (2024). *Lies, Damned Lies, and Distributional Language Statistics: Persuasion and Deception with Large Language Models* (arXiv:2412.17128). arXiv. <https://doi.org/10.48550/arXiv.2412.17128>
- Kran, E., Nguyen, H. M. “Jord,” Kundu, A., Jawhar, S., Park, J., & Jurewicz, M. M. (2025). *DarkBench: Benchmarking Dark Patterns in Large Language Models* (arXiv:2503.10728). arXiv. <https://doi.org/10.48550/arXiv.2503.10728>
- Landemore, H. (Ed.). (2012). *Democratic Reason: Politics, Collective Intelligence, and the Rule of the Many*. Princeton University Press.
- Luettgau, L., Kirk, H. R., Hackenburg, K., Bergs, J., Davidson, H., Ogden, H., Siddarth, D., Huang, S., & Summerfield, C. (2025). *Conversational AI increases political knowledge as effectively as self-directed internet search* (arXiv:2509.05219). arXiv. <https://doi.org/10.48550/arXiv.2509.05219>
- Noggle, R. (2025). *Manipulation: Its Nature, Mechanisms, and Moral Status*. Oxford University Press.
- Röttger, P., Hinck, M., Hofmann, V., Hackenburg, K., Pyatkin, V., Brahman, F., & Hovy, D. (2025a). *IssueBench: Millions of Realistic Prompts for Measuring Issue Bias in LLM Writing Assistance* (arXiv:2502.08395). arXiv. <https://doi.org/10.48550/arXiv.2502.08395>
- Röttger, P., Hinck, M., Hofmann, V., Hackenburg, K., Pyatkin, V., Brahman, F., & Hovy, D. (2025b). *IssueBench: Millions of Realistic Prompts for Measuring Issue Bias in LLM Writing Assistance* (arXiv:2502.08395). arXiv. <https://doi.org/10.48550/arXiv.2502.08395>
- Shiffrin, S. V. (2014). *Speech Matters: On Lying, Morality, and the Law*. Princeton University Press.

- Stanley, J. (2015). *How Propaganda Works*. Princeton University Press.
<https://doi.org/10.2307/j.ctvc773mm>
- Summerfield, C., Argyle, L., Bakker, M., Collins, T., Durmus, E., Eloundou, T., Gabriel, I., Ganguli, D., Hackenburg, K., Hadfield, G., Hewitt, L., Huang, S., Landemore, H., Marchal, N., Ovadya, A., Procaccia, A., Risse, M., Schneier, B., Seger, E., ... Botvinick, M. (2024). *How will advanced AI systems impact democracy?* (arXiv:2409.06729). arXiv.
<https://doi.org/10.48550/arXiv.2409.06729>
- Tessler, M. H., Bakker, M. A., Jarrett, D., Sheahan, H., Chadwick, M. J., Koster, R., Evans, G., Campbell-Gillingham, L., Collins, T., Parkes, D. C., Botvinick, M., & Summerfield, C. (2024). AI can help humans find common ground in democratic deliberation. *Science*, 386(6719), eadq2852. <https://doi.org/10.1126/science.adq2852>
- The White House. (2025). *PREVENTING WOKE AI IN THE FEDERAL GOVERNMENT*.
<https://www.whitehouse.gov/presidential-actions/2025/07/preventing-woke-ai-in-the-federal-government/>

For supplementary materials, please see the supplementary materials section of the arXiv version of the paper here: <https://arxiv.org/abs/2603.10018>