Journal of Experimental Political Science (2026), 1–14
doi:10.1017/XPS.2026.10032

CAMBRIDGE UNIVERSITY PRESS

RESEARCH ARTICLE

A Framework to Assess the Persuasion Risks Large Language Model Chatbots Pose to Democratic Societies

Zhongren Chen[1], Joshua Kalla[1], Quan Le[1], Shinpei Nakamura-Sakai[2], Jasjeet Sekhon[1] and Ruixiao Wang[1]

[1]Yale University, USA and [2]Amazon.com Inc. Seattle, WA, USA
Corresponding author: Joshua Kalla; Email: josh.kalla@yale.edu

Abstract

We investigate whether large language models (LLMs) threaten democracy through their persuasive capabilities. Using two survey experiments ($N = 10{,}417$) and real-world simulations, we compare the cost-effectiveness of LLM chatbots against traditional campaign tactics, taking into account both the "receive" and "accept" steps in the persuasion process. Our design advances prior research by assessing extended human-LLM interactions and measuring short- and long-term effects across three political domains. We find that while LLMs are comparably persuasive to campaign ads once seen, real-world impact depends on both message reception and acceptance. Simulations estimate LLM-based persuasion costs \$48–\$75 per voter versus \$100 for traditional methods. However, traditional methods currently scale more effectively. While LLMs do not yet offer substantially greater potential for large-scale persuasion, this may shift as capabilities improve and techniques for scalable exposure become feasible.

Keywords: political persuasion; large language models; survey experiments; human-AI interaction; political communication

Introduction

In recent years, significant concern has emerged regarding the potential threat that large language models (LLMs) pose to democratic societies through their persuasive capabilities. Numerous scholars, governments, and industry leaders have warned that these systems could undermine democratic processes by manipulating public opinion, amplifying political polarization, and enabling mass-scale computational propaganda (Bengio et al., 2025; Hackenburg and Margetts, 2024; Hackenburg et al., 2025; Goldstein and Sastry, 2023; Goldstein et al., 2023; Shevlane et al., 2023). For example, a malicious foreign actor could use LLMs to generate personalized political

This article has earned badges for transparent research practices: Open Data, Open Materials, and Preregistered. For details see the Data Availability Statement.

© The Author(s), 2026. Published by Cambridge University Press on behalf of American Political Science Association. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

https://doi.org/10.1017/XPS.2026.10032 Published online by Cambridge University Press

content at an unprecedented scale, leading to crises in democratic confidence and civil unrest (Karnofsky, 2024).

In this paper, we propose a framework to assess the persuasive risks that LLM chatbots may pose. Existing frameworks for evaluating the political persuasion impact of LLMs fail to incorporate the real-world challenge of getting potential voters to interact with LLM-produced content. In the real world, political persuasion is a two-step process in which a voter must first *receive* a political message before they can choose whether or not to *accept* that message (Zaller, 1992). There is a well-developed literature on incorporating this stage in assessments of persuasion by campaign messages (Arceneaux and Johnson, 2010; de Benedictis-Kessner et al., 2019) by examining how selective exposure operates in partisan media environments. Existing academic and industry research on LLMs has focused primarily on the second step, answering the question: conditional on being exposed to LLM-produced content in a controlled laboratory setting, how persuasive is an LLM?

We expand upon existing research by conducting two survey experiments and a real-world simulation exercise to determine whether it is more cost-effective to persuade a large number of voters using LLMs compared to standard political campaign practice, taking into account both the receive and accept steps in the persuasion process. We find that, conditional on being exposed to LLM-produced content, LLMs are about as persuasive as actual political campaign TV and digital ads. After taking into account the real-world costs of exposure, we estimate that LLM-based persuasion costs between $48 and $75 per persuaded voter compared to $100 for traditional campaign methods. However, given the difficulties in scalable exposure to LLM-based persuasion, LLM chatbots today likely do not pose a substantial persuasion threat to democratic societies, although this risk is likely to increase with the growing usage of LLMs.

LLMs and political persuasion

Research on LLMs as political persuasion tools has expanded rapidly, yet current approaches might not capture their full real-world impact. First, many existing studies from both academic and industry researchers evaluate LLM persuasiveness using single-message interactions rather than extended conversations with chatbots. For example, prior work has compared LLM-generated persuasive messages to those written by humans but limited their evaluation to single messages (Bai et al., 2025; Hackenburg et al., 2025). This may underestimate LLMs' persuasiveness, as research often finds that extended conversations are particularly persuasive (Broockman and Kalla, 2016; Green and Gerber, 2019; Kalla and Broockman, 2020). LLMs engaging in conversation might be substantially more persuasive as they can promote more active engagement, dynamically respond to voter concerns, and tailor arguments to individual predispositions (Matz et al., 2024; Simchon, Edwards, and Lewandowsky, 2024; Teeny et al., 2021).[1]

Second, when researchers have examined the persuasiveness of LLMs engaged in conversations, they typically do not include a human or campaign standard practice

[1]But see Hackenburg and Margetts (2024) and Argyle et al. (2025) for a conflicting view.

https://doi.org/10.1017/XPS.2026.10032 Published online by Cambridge University Press

comparison (Costello, Pennycook, and Rand, 2024, 2025; Crabtree et al., 2024).[2] Without direct comparisons to human campaign tactics, it remains unclear whether LLMs represent a novel threat.

Third, industry evaluations of LLM persuasive capabilities suffer from methodological flaws that conflate perceived and actual effectiveness. For instance, OpenAI's o1 risk assessment framework classifies political persuasion as a "medium risk" application of their models based primarily on raters' perceptions of a message's persuasiveness rather than experimental evidence of attitude change (OpenAI, 2024). This approach mirrors a common error in political communication research where message effectiveness is judged by perceived persuasiveness rather than demonstrated impact on attitudes or behaviors (O'Keefe, 2018). Given that perceived message effectiveness is rarely correlated with actual message effectiveness (Broockman et al., 2024; O'Keefe, 2018), such evaluations may dramatically over- or under-estimate LLMs' real-world persuasive potential.

Lastly, almost all existing research on LLMs and political persuasion overlooks the first step in the persuasion process: the cost of voter engagement with the persuasive material. Existing studies measure the persuasiveness of LLMs conditional on exposure. However, political campaigns expend substantial resources in the first place simply to get voters to pay attention to their messages through paid advertising, door-to-door canvassing, phone banks, and digital outreach (Limbocker and You, 2020). The challenge of competing for voters' limited attention represents a significant barrier to real-world persuasion that laboratory settings bypass (Barabas and Jerit, 2010; Coppock and Green, 2015; Jerit, Barabas, and Clifford, 2013). Ignoring this step may dramatically overestimate LLMs' threat to democratic processes by implicitly assuming similar exposure costs between human campaign persuasion content and LLMs. In reality, even if LLMs prove exceptionally persuasive when audiences engage with them, their real-world impact might be severely constrained if the costs of achieving widespread engagement remain prohibitively high or if engagement rates with LLM-generated content are substantially lower than with traditional campaign tactics.

Our research addresses each of these limitations. First, we build upon prior research on door-to-door canvassing to test the persuasive impact of LLM chatbots (Broockman and Kalla, 2016; Kalla and Broockman, 2020; Kalla and Broockman, 2023; Kalla, Levine, and Broockman, 2022). Second, we benchmark the persuasiveness of these chatbots against current campaign practices using actual campaign advertisements, providing a meaningful comparison to existing persuasion approaches. Third, we measure both immediate and long-term attitudinal shifts rather than measures of perceived persuasiveness. Finally, we examine not only the second step of the persuasion process (acceptance of the message once exposed) but also the first step (securing voter engagement with LLM content) – allowing us to estimate the real-world threat that LLMs might pose to democratic processes.

[2]This is not intended as a critique of these studies, given that they are interested in different sets of research questions. However, this implies that we cannot extrapolate from the findings in these studies to whether or not LLMs pose a risk to democratic society greater than what humans are currently capable of. And for an exception, see Salvi et al. (2024) and Sehgal et al. (2025).

https://doi.org/10.1017/XPS.2026.10032 Published online by Cambridge University Press

Study 1

We conducted a randomized controlled experiment that compares LLM-driven and traditional human persuasion methods on attitudes toward immigration policy.[3] Participants ($N = 5,150$) recruited from Prolific were asked to complete a survey on the Qualtrics survey platform. They were randomly assigned to one of four experimental conditions:

1. **Placebo condition**: Participants watched a video unrelated to immigration.
2. **Human persuasion**: Participants viewed a video featuring a human advocate presenting pro-immigration arguments.
3. **AI chatbot (as human)**: Participants engaged in interactive conversations with an AI chatbot (using Claude 3.5 Sonnet (20241022-v2)) that presented pro-immigration arguments while identifying itself as a human.
4. **AI chatbot**: Participants engaged in interactive conversations with an AI chatbot (using Claude 3.5 Sonnet (20241022-v2)) that presented pro-immigration arguments while identifying itself as an AI.

For the last two conditions, conversations took place within Qualtrics using a JavaScript-powered chatbox (Molnar, 2019), with the LLM called by the AWS Bedrock API. To maximize the persuasive impact, the prompts relied heavily on the training materials used in prior experiments on the persuasive impact of conversations (Broockman and Kalla, 2016; Kalla and Broockman, 2020; Kalla and Broockman, 2023; Kalla, Levine, and Broockman, 2022; Santoro et al., 2025).[4] The AI was instructed to persuade participants to support the policy position that "illegal immigrants should be eligible for in-state college tuition at state colleges." The prompt engineering included specific guidelines intended to increase its persuasiveness using findings from prior research on persuasive conversations: use an 8th-grade reading level; engage participants in back-and-forth conversation; listen actively and show understanding; build rapport through vulnerability; demonstrate compassion by acknowledging emotions; anticipate and address counter-arguments; and adapt persuasive approaches based on participant responses. The full prompts are presented in Online Appendix Section 7. Median conversations involved 4 participant turns (98.5 words), and 5 chatbot turns (258 words).

Similarly, in an attempt to maximize the persuasive impact of the human persuasion condition, participants watched a video of a teacher sharing their personal reason for supporting this immigration policy. This video was selected due to its large persuasive effect in prior research (Santoro et al., 2025). In this 3-minute video, the confederate shares that he supports the immigration policy because, as a teacher, he had first-hand experience with undocumented students who, despite their academic excellence, were unable to afford to attend college due to their immigration status.

[3]This policy was chosen because it was previously used in prior studies (Kalla and Broockman, 2020; Santoro et al., 2025).

[4]We conducted four pilot studies with multiple prompts where we read conversation transcripts to determine whether it seemed like the chatbots were following instructions and being persuasive.

https://doi.org/10.1017/XPS.2026.10032 Published online by Cambridge University Press

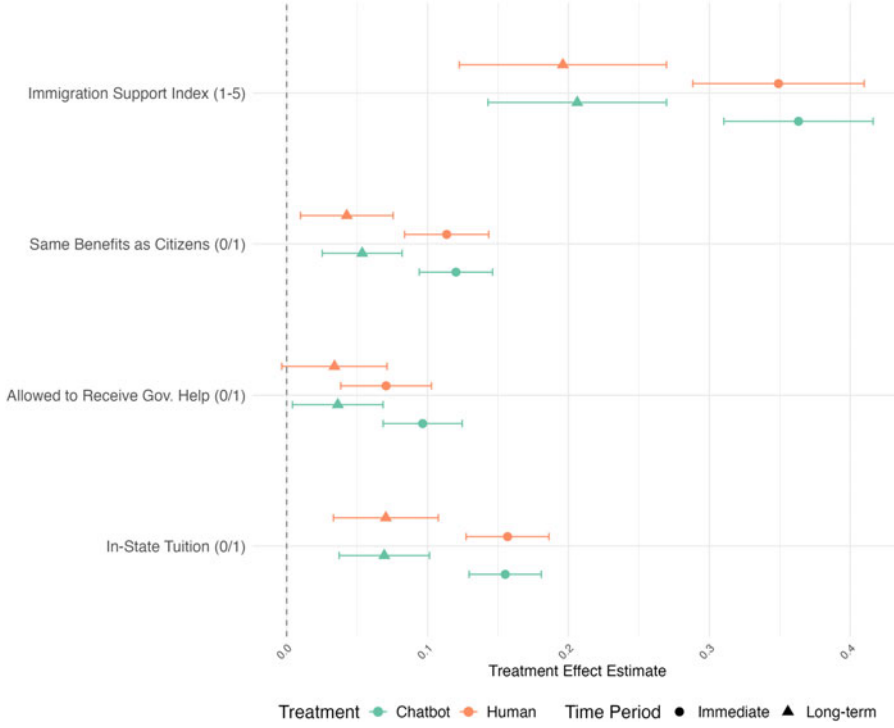


Figure 1. 95% confidence interval on the average treatment effect compared to placebo in the short term (post-survey) and long term (5 weeks after the survey). See Table A.5 for full results.

We measured outcomes immediately posttreatment and five weeks later using three policy items, rated on five-point Likert agreement scales: whether undocumented immigrants should be eligible for in-state college tuition at state colleges, allowed to receive government help to pay for college, and receive the same government benefits as American citizens. To reduce measurement error, we compute an immigration support index by averaging the three policy items. We also re-code the individual items as binary measures, where 1 is any support for the pro-immigration policy, and 0 is any opposition or indifference. $N = 3,412$ participants completed the follow-up survey five weeks later. Details on tests of covariate balance and differential attrition are in Tables A.2–A.4. To estimate average treatment effects, we use linear regression with preregistered pretreatment covariates to increase precision.

The preregistration document is available at https://osf.io/s976d/files/4vrpb?view_only = 1b4aff47cf744391beafe3ba3dccdf78. See Additional Supplementary Materials for information about the correspondence between the manuscript and the preregistration document. This study was conducted in accordance with the APSA Principles and Guidance for Human Subjects Research.

Figure 1 shows our main results. We combine the two AI conditions into a single group. Across both our additive index and each outcome measure, we find that both the chatbot and human persuasion conditions are persuasive relative to the placebo

https://doi.org/10.1017/XPS.2026.10032 Published online by Cambridge University Press

immediately posttreatment and in our five-week follow-up survey. Relative to the placebo, the chatbot shows an effect of 0.363 scale points ($SE = 0.027$) immediately and 0.206 scale points ($SE = 0.032$) after five weeks. The corresponding effects for human conditions are 0.349 scale points ($SE = 0.031$) and 0.196 scale points ($SE = 0.038$). However, there is no distinguishable difference between the chatbot conditions and the human persuasion condition either immediately posttreatment ($p = 0.601$) or in the five-week follow-up survey ($p = 0.758$). Figure 1 presents additional results separately by each policy item in the additive index, similarly finding no distinguishable difference. Table A.5 contains full numerical point estimates, standard errors, and p-values. Table A.6 presents the estimated average treatment effects comparing AIs that identify as human versus those that identify as AI. The results show no significant difference in persuasive effectiveness between the two AI conditions, consistent with other research findings that authorship labels have minimal effect on the persuasiveness of AI-generated content (Boissin et al., 2025; Gallegos et al., 2025; Goel et al., 2025). The estimated effect of AI identifying as human, relative to AI identifying as AI, is 0.00 scale points ($SE = 0.033$) immediately and 0.018 scale points ($SE = 0.038$) after five weeks. Online Appendix Section 1.9 examines the association between the number of conversation turns and the persuasive effect. We find a positive association between the number of turns and the persuasive outcome. Table A.7 reports a non-preregistered equivalence test (Rainey, 2014) of AI versus Human.

Taken together, these findings suggest that AI chatbots can be as persuasive as human persuaders, both in the short-term and up to five weeks later. When it comes to the second step of persuasion, there is no significant difference between the two. Before comparing cost-effectiveness, we note several Study 1 limitations that we address in a second study. First, the human persuasion condition was a 3-minute video, substantially longer than the typical 30 or 60 second TV and digital ads used by political campaigns. This longer video could overstate the effectiveness of human persuasion. Second, this study only investigated one political issue on immigration. The relative effectiveness of AI chatbots compared to human persuasion might vary across different domains. Third, LLMs continue to advance; more recent models might be more persuasive.

Study 2

We conducted a second survey experiment on Prolific ($N = 5,267$) on three issues: immigration, transgender rights, and minimum wage. The persuasion treatments were designed to encourage participants to support the following policy positions:

1. Illegal immigrants should be eligible for in-state college tuition at state colleges.
2. Transgender people should be allowed to use the restroom that matches the gender they live every day.
3. The federal minimum wage should not be increased from the current $7.25 per hour to $15 per hour.

https://doi.org/10.1017/XPS.2026.10032 Published online by Cambridge University Press

The third policy ensures testing persuasion in both liberal and conservative directions. This study was conducted on the Qualtrics survey platform in a manner similar to Study 1. Participants were randomly assigned to a policy, and then within each policy, participants were randomly assigned to a placebo, a human persuasion, or an AI chatbot condition. The human persuasion condition involved watching a 30–60 second campaign advertisement, a scenario closely aligned with real-world political ads. The AI chatbot condition used in this study employed similar prompts to those in Study 1 and was powered by Claude 3.7 Sonnet, which outperforms Claude 3.5 Sonnet in advanced reasoning (Anthropic, 2025). The full prompts are presented in Online Appendix Section 7. Median conversation involved 4 participant turns (89 words), and 5 chatbot turns (281 words). Tables A.12 and A.13 present tests of covariate balance and differential attrition.

We measured outcomes immediately posttreatment using a single issue-specific policy item rated on five-point Likert agreement scale: whether illegal immigrants should be eligible for in-state college tuition at state colleges; whether transgender people should be allowed to use the restroom that matches the gender they live every day; or whether the federal minimum wage should be increased from the current $7.25 per hour to $15 per hour. We also re-code the individual items as binary measures, where 1 indicates any support for the policy position advocated in the treatment, and 0 represents either opposition or indifference. To estimate average treatment effects, we use linear regression with preregistered pretreatment covariates.

The preregistration document is available at https://osf.io/s976d/files/hd43t?view_only = 1b4aff47cf744391beafe3ba3dccdf78. See Additional Supplementary Materials for information about the correspondence between the manuscript and the preregistration document. This study was conducted in accordance with the APSA Principles and Guidance for Human Subjects Research.

Figure 2 shows the results. Table A.14 contains full numerical point estimates and standard errors for each outcome. Table A.15 presents the point estimates, standard errors, and p-values comparing the AI persuasion and human persuasion, pooled across the three issues. Across all issues, we find no consistent evidence that the AI chatbot is more persuasive than human persuasion. If anything, when pooling across the three issues, the human persuasion condition shows slightly larger persuasive effects than the AI chatbot. We estimate that the human persuasion is 0.001 percentage points more persuasive than the chatbot ($SE = 0.010$; $p = 0.896$) for the pooled binary outcome and 0.059 scale points more persuasive (0.025; $p = 0.018$) for the pooled continuous outcome. Robustness check among participants who, in the pretreatment survey, opposed the position advocated in the treatment condition, yields similar effects (Tables A.17–A.18). Online Appendix Section 2.10 examines the association between the number of conversation turns and the persuasive effect for each issue. We find a positive association between the number of turns and the persuasive outcome. Table A.16 reports a non-preregistered equivalence test (Rainey, 2014) of AI versus Human. Online Appendix Section 8 presents a non-preregistered investigation of the on-topic engagement rates and AI persuasion strategies in Studies 1 and 2.

To conclude, even when (1) the human persuasion condition is limited to a 30–60 second campaign video, (2) a more powerful language model – Claude 3.7

https://doi.org/10.1017/XPS.2026.10032 Published online by Cambridge University Press

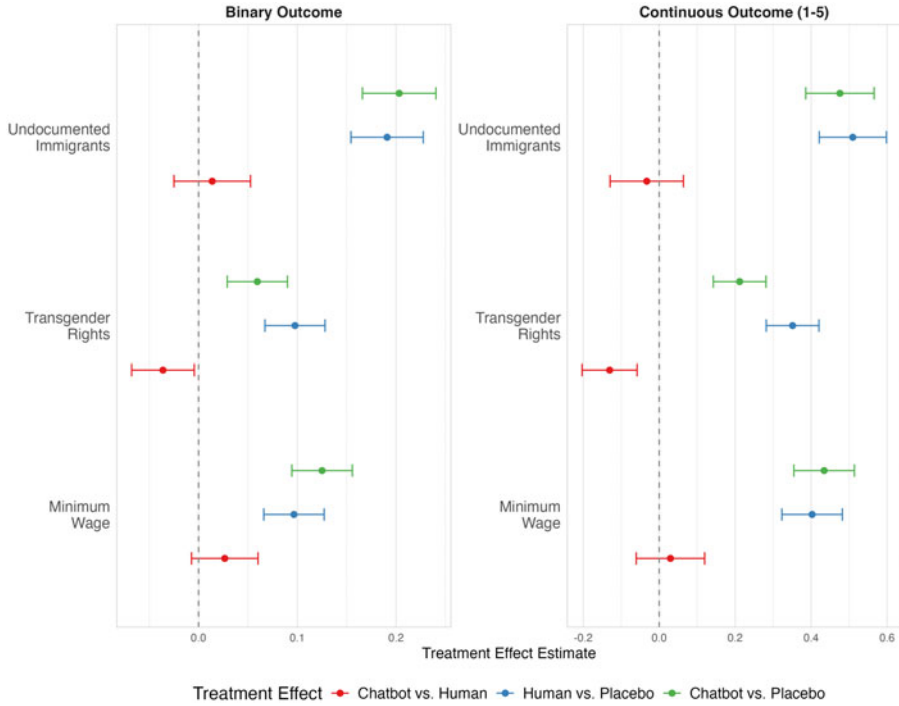


Figure 2. 95% confidence interval on the average treatment effect of AI versus human and placebo on three different topics. Tables A.14–15 contain full results.

Sonnet – is used, and (3) the persuasion topics span both sides of partisan issues, we find no evidence that AI chatbots outperform human persuasion.

Real-world cost-efficacy analysis

While both AI-based and standard campaign persuasion efforts were comparably effective in shifting policy attitudes under conditions of forced exposure – for example, in Study 1, both conditions led to a $\approx$13-percentage-point increase in support for the immigration policy, and a $\approx$5-point increase one month later – real-world persuasion depends on securing exposure, a key often overlooked cost. In this section, we present the results of a bounding exercise to estimate the relative cost efficacy in the real world.

First, our human persuasion estimates from the survey experiment need to be adjusted. In our survey experiments, subjects were forced to watch the human persuasion video. They were also instructed to pay careful attention to the video. Both of these features likely overstate the real-world efficacy of our human persuasion. To adjust for this, we conducted a survey of 12 academics and campaign practitioners who are experts in survey experiments and their implications for real-world campaign effects (17 were invited, for a 71% response rate). We asked these experts their priors for how our survey experimental evidence would generalize to

https://doi.org/10.1017/XPS.2026.10032 Published online by Cambridge University Press

the real-world, asking specifically about non-skippable YouTube video ads (see Online Appendix Section 4 for full question wording).[5] The median expert estimated that the real-world treatment effect would be roughly 14% the size of the survey experiment treatment effect, with a minimum estimate of 1% and a maximum estimate of 60%. Given the often minimal effects of real-world persuasion (Kalla and Broockman, 2018) and experts' overconfidence in estimating persuasiveness (Broockman et al., 2024), for the purposes of this exercise, we conservatively assume that the real-world efficacy would be 1% that of our survey experiment estimate, using the smallest deflator selected by an expert. As a result, our best guess is that over the long term, the human persuasion would produce a 0.05 percentage-point treatment effect (5 * 0.01).

To estimate the cost-effectiveness of such persuasion at scale, we draw on two independent cost estimates. Based on data from Google, non-skippable YouTube ads cost between $2.70 and $8.10 per 1,000 impressions. Political practitioners we consulted estimated a substantially higher real-world average of $50 per 1,000 impression given recent costs from the 2024 election cycle.[6] As a conservative estimate of the cost-effectiveness of human persuasion, using the higher cost estimate from the 2024 election cycle, we estimate a cost per persuaded voter of $100 (50/1000/.0005). Additionally, YouTube has a large inventory of non-skippable advertisements. For example, YouTube estimates that a month-long advertising campaign would reach an estimated 42 million unique viewers, yielding an estimated 21,000 net new votes.

Whereas human persuasion is quite scalable and cost-effective, AI persuasion faces a fundamental difficulty: the requirement for voluntary opt-in by the target audience to speak with a chatbot. To test the scalability and expense required to persuade a large number of people, we experimented with three different recruitment methods for AI interaction.

First, we launched an advertising campaign on Meta encouraging people to have a conversation with a chatbot (see Figure 3a). We spent $199.74 on ads, which produced 53 conversations where the user made at least one statement to the AI chatbot. Inference cost around $0.05 per conversation. If we assume the treatment effect from these conversations is as large as those in our survey experiments, this implies a cost per conversation incorporating advertising and inference costs of $3.77 and a cost per vote of $75.

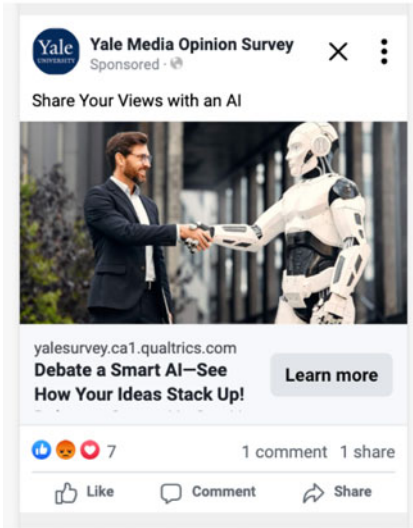
Second, we launched a follow-up advertising campaign on Meta where we offered to pay participants $1 for having a conversation with a chatbot (see Figure 3b). We spent $96.35 on ads and $168 on gift cards to produce 102 conversations where the user made at least one statement to the LLM chatbot and 73 where the user made at least two. If we assume the treatment effect from these conversations is as large as

[5]Non-skippable YouTube video ads are comparable to television advertising in how they operate. Non-skippable ads need to be watched in full before the desired content is shown, similar to how television ads operate. YouTube ads were chosen for this exercise because, unlike TV ads, there is greater cost transparency.

[6]Ad costs tend to increase closer to a major election, given increased demand from multiple political campaigns.

https://doi.org/10.1017/XPS.2026.10032 Published online by Cambridge University Press

(a) Example Uncompensated Meta Ad



(b) Example Compensated Meta Ad

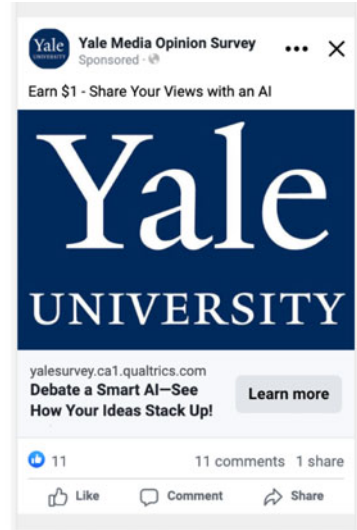


Figure 3. Meta advertisements.

those in our survey experiments, this implies a cost per one-statement conversation of $2.59 and a cost per vote of $52.

Lastly, on Prolific, a campaign can get voters to have a conversation with an AI chatbot for around $2.39, between participant payment, Prolific's fees, and inference costs. Assuming a 5 percentage-point treatment effect, this implies a cost per net vote of $48.

However, it could be difficult to scale these outreach methods. Prolific has around 110,000 active users in the United States. Reaching all of these users with an LLM chatbot would yield an estimated 5,500 net new votes. On Meta, 8,253 users saw the paid advertisement, which produced 73 conversations where the user made at least two statements to the LLM chatbot. Reaching 130 million American adults (roughly half of all American adults) with the Meta ad would yield an estimated 1.1 million two-statement conversations and 57,000 votes.[7]

Across these three different recruitment modes, we estimate that AI-based persuasion costs between $48 and $75 per vote, potentially less expensive than the $100 per vote we estimate for human persuasion. However, it could be challenging to scale the LLM-based persuasion given the limited size of the Prolific audience and the low conversation conversion rate on Meta. These results suggest that while AI-based persuasion can match human performance on a per-person basis under ideal conditions of forced exposure, its real-world deployment is currently constrained by exposure costs and audience sizes.

[7]Potentially, this is an upper bound for the number of votes that could be produced via Meta ads, given that these ads were run by a university.

https://doi.org/10.1017/XPS.2026.10032 Published online by Cambridge University Press

Importantly, this exercise has several important limitations. We assume a conservative estimate of the real-world efficacy of human persuasion, taking the minimum estimate from our expert survey. On the other hand, we assume the real-world efficacy of our chatbot would be 100% that of our survey experiment. If the subjects in our survey experiment are more persuadable or more willing to engage with the chatbot than the typical American (which we find to be the case when comparing the transcripts of chats collected via Prolific compared to Meta), we could be overstating the efficacy of chatbot-based persuasion. Furthermore, our framework treats the *receive* and *accept* stages as independent events. But in reality, a voter's susceptibility to persuasion is almost certainly correlated with their willingness to engage with the message. Future research should consider better real-world estimates of these treatment effects.

Conclusion

Our research introduces a new framework for assessing the persuasive risks that LLM chatbots pose to democratic societies. By conducting two survey experiments and a real-world simulation exercise, we evaluated both steps of the persuasion process: exposure to a persuasive message and acceptance conditional on exposure. While LLMs match human persuasion under forced exposure, their real-world impact is constrained by scale.

Through our cost-efficacy analysis, we estimate that LLM-based persuasion costs between $48 and $75 per persuaded voter compared to approximately $100 for traditional campaign methods. However, the scalability limitations of LLM-based approaches – including limited audience pools and low conversation conversion rates – present potential barriers to widespread political influence. Our findings suggest that although AI chatbots can match human performance under forced exposure, the practical difficulties in achieving exposure at scale might currently limit their threat to democratic processes.

This study has several important limitations. First, we only study the persuasive impact of chatbots. This was motivated by their widespread adoption and prior research finding large persuasive impacts from conversations (Broockman and Kalla, 2016; Green and Gerber, 2019; Kalla and Broockman, 2020, 2023). However, other forms of AI persuasion, such as flooding social media, may prove to be more cost-effective. Furthermore, we only studied the persuasive impact of a single chatbot conversation. Chatbots will very likely become more persuasive over time as users come to trust them and form parasocial relationships with them (Maeda and Quan-Haase, 2024). However, the costs of such long-term use would also be high. Second, we only studied three political issues. It is possible that on other issues or with different prompts, the AI may perform quite differently. However, we investigated the persuasive impacts of AI on three distinct economic and cultural issues, finding that regardless of the issue, human and AI persuasion have similar effects. Third, new AI models are rapidly being released. We only investigated the impact of two models. Fourth, participants were informed ex ante that they would engage with an AI in the context of a scientific study, which may have increased the AI's perceived neutrality. In practice, skepticism toward AI may attenuate

https://doi.org/10.1017/XPS.2026.10032 Published online by Cambridge University Press

persuasive effects, warranting further research on how source cues affect persuasion. Fifth, we compare an active chatbot treatment to a passive human condition where a participant watches a video. Online Appendix Section 3 presents an exercise comparing our chatbot results to the door-to-door canvassing results in Kalla and Broockman (2020). Finally, while the present approach assumes independence between the *receive* and *accept* stages, the probability of a voter accepting a message is likely conditional on their initial propensity to receive it. Our cost-efficacy analysis remains observational. Future work should employ experimental joint choice designs to estimate these dynamics.

As LLM capabilities improve and user engagement with chatbots grows, ongoing assessment will be essential. For now, our research indicates that LLMs do not currently pose a substantially greater threat to democratic societies through mass persuasion than existing human-driven methods. However, this risk may change as LLM adoption grows. It is also crucial to recognize that the AI persuasion capability can be beneficial by democratizing political communication for under-resourced advocacy organizations, lowering the financial barriers to civic engagement, and facilitating prejudice reduction (Costello et al., 2025; Crabtree et al., 2024). As such, our framework provides a tool to evaluate both the potential threats and benefits to democratic societies.

Supplementary material. The supplementary material for this article can be found at https://doi.org/10.1017/XPS.2026.10032.

Data availability. Upon publication, replication data and code from all studies will be made available at Chen et al. (2026).

Acknowledgements. We thank Ella Barrett, Joe Benton, Stephen Deline, Brian Goodrich, Jared Kaplan, Holden Karnofsky, Max Nadeau, David Rand, Otis Reid, Jonathan Robinson, Josh Rosmarin, Semra Sevi, David Shor, Aaron Strauss, Christopher Summerfield, Jan Voelkel, and Michelle Zeiler, seminar participants at Amazon, and attendees of the Experimental Designs in the Era of Artificial Intelligence Workshop at Berkeley for helpful comments. All remaining errors are our own. Author names are listed in alphabetical order.

Competing interests. The authors declare no conflicts of interest.

Funding. This research was funded by a grant from Coefficient Giving.

Ethics statement. This research was reviewed and deemed exempt by the Yale University Human Subjects Committee (Yale IRB Protocol #2000038320). This research adhered to the APSA Principles and Guidance for Human Subjects Research.

References

Anthropic. 2025. Claude 3.7 Sonnet and Claude Code. https://www.anthropic.com/news/claude-3-7-sonnet (March 23, 2025).

Arceneaux, Kevin, and Martin Johnson. 2010. "Does Media Fragmentation Produce Mass Polarization? Selective *Exposure and a New Era of Minimal Effects*." *Selective Exposure and a New Era of Minimal Effects*.

Argyle, Lisa P, Ethan C. Busby, Joshua R. Gubler, Austin Lyman, Jeremy Olcott, Julianne Pond, and David Wingate. 2025. "Testing Theories of Political Persuasion Using AI." *Proceedings of the National Academy of Sciences 122*(18): e2412815122.

Bai, Hui, Jan G. Voelkel, Shane Muldowney, C. Eichstaedt Johannes, and Robb Willer. 2025. "AI-Generated Messages Can Be Used to Persuade Humans on Policy Issues." *OSF Preprints*.

Barabas, Jason, and Jennifer Jerit. 2010. "Are Survey Experiments Externally Valid?" *American Political Science Review 104*(2): 226–42.

Bengio, Yoshua, Sören Mindermann, Daniel Privitera, Tamay Besiroglu, Rishi Bommasani, Stephen Casper, Yejin Choi, Philip Fox, Ben Garfinkel, and Danielle Goldfarb. 2025. "International AI Safety Report." *arXiv [Preprint]*.

Boissin, Esther, Thomas H. Costello, Daniel M. Alonso, David G. Rand, and Gordon Pennycook. 2025. "AI Reduces Conspiracy Beliefs Even When Presented as a Human Expert."

Broockman, David E, Joshua L. Kalla, Christian Caballero, and Matthew Easton. 2024. "Political Practitioners Poorly Predict Which Messages Persuade the Public." *Proceedings of the National Academy of Sciences 121*(45): e2400076121.

Broockman, David, and Joshua Kalla. 2016. "Durably Reducing Transphobia: A Field Experiment on Door-to-Door Canvassing." *Science 352*(6282): 220–4.

Chen, Zhongren, Joshua Kalla, Quan Le, Shinpei Nakamura-Sakai, Jasjeet Sekhon, and Ruixiao Wang. 2026. "Replication Data for: A Framework to Assess the Persuasion Risks Large Language Model Chatbots Pose to Democratic Societies." https://doi.org/10.7910/DVN/TGKINF.

Coppock, Alexander, and Donald P. Green. 2015. "Assessing the Correspondence Between Experimental Results Obtained in the Lab and Field: A Review of Recent Social Science Research." *Political Science Research and Methods 3*(1): 113–31.

Costello, Thomas H, Gordon Pennycook, and David G. Rand. 2024. "Durably Reducing Conspiracy Beliefs through Dialogues with AI." *Science 385*(6714): eadq1814.

Costello, Thomas H, Gordon Pennycook, and David G. Rand. 2025. "Just the Facts: How Dialogues with AI Reduce Conspiracy Beliefs." *OSF [Preprint]*.

Costello, Thomas H, Gordon Pennycook, Robb Willer, and David G. Rand. 2025. "Deep Canvassing Using AI."

Crabtree, Charles, John Holbein, Mitchell Bosley, and Semra Sevi. 2024. "Can AI Reduce Prejudice at Scale? Evaluating the Effectiveness of AI-Powered Personalized Persuasion on Support for Transgender Rights." *Preprint*.

de Benedictis-Kessner, Justin, Matthew A. Baum, Adam J. Berinsky, and Teppei Yamamoto. 2019. "Persuading the Enemy: Estimating the Persuasive Effects of Partisan Media with the Preference-Incorporating Choice and Assignment Design." *American Political Science Review 113*(4): 902–16.

Gallegos, Isabel O., Chen Shani, Weiyan Shi, Federico Bianchi, Izzy Gainsburg, Dan Jurafsky, and Robb Willer. 2025. "Labeling Messages as AI-Generated Does Not Reduce Their Persuasive Effects." *arXiv Working Paper*.

Goel, Natasha, Thomas Bergeron, Blake Lee-Whiting, Danielle Bohonos, Md Mujahedul Islam, Sarah Lachance, Sonja Savolainen, Clareta Treger, and Eric Merkley. 2025. "Artificial Influence: Comparing the Effects of AI and Human Source Cues in Reducing Certainty in False Beliefs." *OSF Working Paper*.

Goldstein, Josh A, and Girish Sastry. 2023. "The Coming Age of AI-powered Propaganda. How to Defend Against Supercharged Disinformation." *Foreign Affairs*.

Goldstein, Josh A, Girish Sastry, Micah Musser, Renee DiResta, Matthew Gentzel, and Katerina Sedova. 2023. "Generative Language Models and Automated Influence Operations: Emerging Threats and Potential Mitigations." *arXiv [Preprint]*.

Green, Donald P, and Alan S. Gerber. 2019. *Get Out the Vote: How to Increase Voter Turnout*. Washington, D.C.: Brookings Institution Press.

Hackenburg, Kobi, Ben M. Tappin, Paul Röttger, Scott A. Hale, Jonathan Bright, and Helen Margetts. 2025. "Scaling Language Model Size Yields Diminishing Returns for Single-Message Political Persuasion." *Proceedings of the National Academy of Sciences 122*(10): e2413443122.

Hackenburg, Kobi, 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.

Jerit, Jennifer, Jason Barabas, and Scott Clifford. 2013. "Comparing Contemporaneous Laboratory and Field Experiments on Media Effects." *Public Opinion Quarterly 77*(1): 256–82.

Kalla, Joshua L, Adam Seth Levine, and David E. Broockman. 2022. "Personalizing Moral Reframing in Interpersonal Conversation: A Field Experiment." *The Journal of Politics 84*(2): 1239–43.

https://doi.org/10.1017/XPS.2026.10032 Published online by Cambridge University Press

Kalla, Joshua L, and David E. Broockman. 2018. "The Minimal Persuasive Effects of Campaign Contact in General Elections: Evidence from 49 Field Experiments." *American Political Science Review 112*(1): 148–66.

Kalla, Joshua L, and David E. Broockman. 2020. "Reducing Exclusionary Attitudes through Interpersonal Conversation: Evidence from Three Field Experiments." *American Political Science Review 114*(2): 410–25.

Kalla, Joshua L, and David E. Broockman. 2023. "Which Narrative Strategies Durably Reduce Prejudice? Evidence from Field and Survey Experiments Supporting the Efficacy of Perspective-Getting." *American Journal of Political Science 67*(1): 185–204.

Karnofsky, Holden. 2024. "A Sketch of Potential Tripwire Capabilities for AI." *Carnegie Endowment for International Peace* (March 10, 2025).

Limbocker, Scott, and Hye Young You. 2020. "Campaign Styles: Persistency in Campaign Resource Allocation." *Electoral Studies 65*: 102140.

Maeda, Takuya, and Anabel Quan-Haase. 2024. "When Human-AI Interactions become Parasocial: Agency and Anthropomorphism in Affective Design." *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency* 1068–1077.

Matz, Sandra C, Jacob D. Teeny, Sumer S. Vaid, Heinrich Peters, Gabriella M. Harari, and Moran Cerf. 2024. "The Potential of Generative AI for Personalized Persuasion at Scale." *Scientific Reports 14*(1): 4692.

Molnar, Andras. 2019. "SMARTRIQS: A Simple Method Allowing Real-Time Respondent Interaction in Qualtrics Surveys." *Journal of Behavioral and Experimental Finance 22*: 161–9.

O'Keefe, Daniel J. 2018. "Message Pretesting Using Assessments of Expected or Perceived Persuasiveness: Evidence About Diagnosticity of Relative Actual Persuasiveness." *Journal of Communication 68*(1): 120–42.

OpenAI. 2024. OpenAI o1 System Card. https://arxiv.org/abs/2412.16720.

Rainey, Carlisle. 2014. "Arguing for a Negligible Effect." *American Journal of Political Science 58*(4): 1083–91.

Salvi, Francesco, Manoel Horta Ribeiro, Riccardo Gallotti, and Robert West. 2024. "On the Conversational Persuasiveness of Large Language Models: A Randomized Controlled Trial." arXiv [Preprint].

Santoro, Erik, David E. Broockman, Joshua L. Kalla, and Roni Porat. 2025. "Listen for a Change? A Longitudinal Field Experiment on Listening's Potential to Enhance Persuasion." *Proceedings of the National Academy of Sciences 122*(8): e2421982122.

Sehgal, Neil K. R, Sunny Rai, Manuel Tonneau, Anish K. Agarwal, Joseph Cappella, Melanie Kornides, Lyle Ungar, Alison Buttenheim, and Sharath Chandra Guntuku. 2025. "Conversations with AI Chatbots Increase Short-Term Vaccine Intentions But Do Not Outperform Standard Public Health Messaging." *arXiv*.

Shevlane, Toby, Sebastian Farquhar, Ben Garfinkel, Mary Phuong, Jess Whittlestone, Jade Leung, Daniel Kokotajlo, Nahema Marchal, Markus Anderljung, and Kolt Noam. 2023. " Model Evaluation for Extreme Risks." *arXiv [Preprint]*.

Simchon, Almog, Matthew Edwards, and Stephan Lewandowsky. 2024. "The Persuasive Effects of Political Microtargeting in the Age of Generative Artificial Intelligence." *PNAS Nexus 3*(2): pgae035.

Teeny, Jacob D, Joseph J. Siev, Pablo Briñol, and Richard E. Petty. 2021. "A Review and Conceptual Framework for Understanding Personalized Matching Effects in Persuasion." *Journal of Consumer Psychology 31*(2): 382–414.

Zaller, John. 1992. *The Nature and Origins of Mass Opinion*. Cambridge: Cambridge University Press.

Cite this article: Chen Z, Kalla J, Le Q, Nakamura-Sakai S, Sekhon J, and Wang R (2026). A Framework to Assess the Persuasion Risks Large Language Model Chatbots Pose to Democratic Societies. *Journal of Experimental Political Science*. https://doi.org/10.1017/XPS.2026.10032

https://doi.org/10.1017/XPS.2026.10032 Published online by Cambridge University Press