nature portfolio

Peer Review File

Persuading Voters using Human-AI Dialogues

Corresponding Author: Professor David G. Rand

Any redactions in this file are there to maintain patient confidentiality, the confidentiality of unpublished data, or to remove third-party material.

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Version 2:

Reviewer comments:

Referee #1

(Remarks to the Author)
A. Key Results
The manuscript provides an empirical examination of whether (and how) large language models (LLMs) can persuade voters in the context of the 2024 U.S. Presidential Election and a Massachusetts state ballot measure on legalizing psychedelics. Notable findings include:

*Measurable Persuasion Effects: Brief, dialog-based interactions with AI significantly shifted participants' candidate preference and voting intentions—often by a few points on a 0-100 scale, sometimes comparable or exceeding shifts seen in traditional campaign ads. Around one-third of the initial effect persisted over a month later, suggesting that AI persuasion effects are not entirely short-lived.

*Asymmetry in Accuracy: The AI advocating for Trump was more prone to misstatements and inaccuracies compared to the AI advocating for Harris, which generally provided more accurate information.

*Versatility of AI: The dialogues were persuasive for both highly polarized and less-polarized contexts—namely, presidential candidates and a ballot measure on psychedelics.

*Likely Mechanism: The manuscript presents correlational evidence that politeness, fact-based arguments, and personalized responses—rather than negativity or anger inducement — were associated with more persuasion.

Together, these features highlight the potential for AI-driven political persuasion, raise ethical considerations about misinformation, and provide data-driven insights into AI's capabilities as a political influencer.

B. Originality and Significance
The paper offers a novel contribution by empirically testing human–AI dialogue (rather than static AI-generated text) as a mechanism for political persuasion. While a rich literature exists on minimal campaign effects, this study extends that conversation by showing larger-than-expected persuasion, particularly via short, personalized dialogues.

Originality:
*Few prior works have systematically measured real-time conversation-based AI persuasion in a controlled manner.

* While the study addresses a timely question—whether AI can effectively persuade voters—the depth of new insight for psychological science is pretty limited. Despite employing a new medium (chatbots), the core takeaways—"people can be persuaded by messages consistent with their prior beliefs or that align with core concerns"—are neither surprising nor substantially more advanced than existing research on persuasion. For instance, work over the last several decades demonstrates the benefits of personalizing messages to recipients in ways that are aligned with their values, identity, or personal disposition (e.g. Cialdini, 2001; Feingberg & Willer 2013, 2015; Petty & Cacioppo, 1984; Wheeler, Petty, & Bizer, 2005). Additional studies further demonstrate the benefits of more positively-oriented and polite messages for persuading audiences (Broockman & Kalla, 2016; Matthes & Marquart, 2015).

Significance:
The results are timely, given the rapid adoption of large language models. These findings will interest political scientists, psychologists of persuasion, campaign strategists, and policy-makers concerned with regulating AI in electoral contexts.

*Though the authors used multiple agent systems (GPT-4o, Perplexity, etc.), the approach is still a bit narrow. For example, the study acknowledges that the AI occasionally produced inaccuracies, but some LLMs may be more prone to misinformation than others. A thorough cross-LLM comparison or systematic manipulation of the guardrails seems important if the goal is to generalize to LLMs as a class of technology (versus positioning the results as only pertaining to certain models) to inform a general audience.

*The persuasive instructions were carefully crafted by the authors, which is somewhat contrived compared to how political operatives might attempt to push an AI's limits. A stronger contribution would include evidence of generalizability across multiple prompt designs, more open-ended dialogues, or widely different LLMs.

C. Data & Methodology

Strengths:

*The authors used a multi-agent system to generate real-time answers. This is both creative and relevant, ensuring up-to-date responses while still controlling the conversation topic.

*The pre-/post- design, large sample sizes, and random assignment to conditions support internal validity.

*The authors compare AI persuasion impacts with known benchmarks (e.g., standard campaign advertising), adding external context.

*Sampling & Randomization: The study recruited over 2,000 participants for the presidential setting and 500 for the ballot measure; randomization appears thorough, with no major imbalances. The manuscript describes how dropouts and pre-treatment covariates were examined, with no major imbalances detected.

*Transparency & Reproducibility: The method sections—especially with references to the pre-registrations—are detailed. The paper also references robust extended data (e.g., randomization scripts, conversation transcripts) and invites readers to see further materials in the supplement.

Concerns:

*Selection or Opt-In Bias: Participants were told that they were interacting with an AI agent; someone who chooses to spend several minutes conversing with a chatbot may be more open to new arguments—especially if they're already on a survey platform and expecting to complete a task. In actual campaigns, real-life distractions, cynicism, or reluctance might yield lower persuasion.

*Generality Across Different Populations: While the authors recruited from reputable online panels, differences in digital literacy, political interest, or demographic composition might affect generalizability. Some discussion or additional data on how representative these samples are of U.S. voters would be important.

*Reliance on correlational data for insights: One of the key contributions of the paper potentially is the insight on the strategies that are most persuasive. However, the paper does not do enough experimental manipulation of the conversation strategies themselves. Simply observing "the AI used X strategy" is correlational.

*Unclear Mechanisms: There is no firm causal demonstration of why certain strategies succeed (or why accuracy does not correlate with persuasive success). For a top-tier journal, deeper theoretical integration—tying these findings to established frameworks of persuasion, psychology, and computational communication—is typically expected. A fully randomized design of distinct strategy prompts—plus a carefully measured causal analysis of which strategies truly cause persuasion—would make the paper more mechanistically robust and worthy of high-impact publication.

D. Appropriate Use of Statistics and Treatment of Uncertainties

*Statistical Tests: The authors primarily use OLS regressions with heteroscedasticity-robust standard errors, difference-in-means tests, and follow-up cross-validation (for persuasion strategy features).

*Figures and regression tables include 95% confidence intervals; these appear clearly labeled. The paper consistently reports relevant estimates—e.g., effect size changes of 2–3 points on a 0–100 scale. This is commendable because effect sizes are more meaningful than p-values alone.

*Potential Multiple Comparisons: The authors explore multiple outcomes with no clear prediction on direction of change

(candidate preference, vote intention, etc.). While they note the standard disclaimers, they should apply a multiple-testing correction to adjust the effect size and reassure readers that the strong results are not false positives.

*A concern is that although the authors compare their effect sizes to typical campaign ads, some might question whether a 2–3 point shift on a 0–100 scale translates into a meaningful real-world impact (assuming that is the magnitude after adjustment for post-hoc comparisons). Converting this small numerical shift into claims of "meaningful influence on election outcomes" could be an overreach—particularly in high-salience national elections where many other competing messages exist.

*The study finds that around a third of the effect persists after a month, but no direct assessment of real behavior (e.g., actual turnout or donation patterns) was done.

E. Conclusions

*Robustness: The immediate persuasion effect is clear, and the persistence analysis indicates that at least one-third of the effect remains over time. It is essential to ensure appropriate adjustments for post-hoc comparisons are made before interpreting the magnitude of these effects.

*Interpretation: The authors acknowledge that although AI dialogues can be persuasive, real-world usage may face challenges (e.g., getting people to opt in). They avoid exaggeration and place their results in the context of existing minimal-effects literature.

F. Suggested Improvements

*Test robustness across different LLMs: Though the authors used multiple agent systems (GPT-4o, Perplexity, etc.), the approach might still seem narrow. If the intent is to generalize to this class of AI models, a thorough cross-LLM comparison or systematic manipulation of the guardrails seems important

*Conduct experiments to test causal mechanisms: There is no firm causal demonstration of why certain strategies succeed (or why accuracy does not correlate with persuasive success). A fully randomized design of distinct strategy prompts—plus a carefully measured causal analysis of which strategies truly cause persuasion—would make the paper more mechanistically robust and a more valuable contribution to the scientific literature.

*Broader Demographic Representation: Additional analysis of sample versus population or expanding sample to enable demographic breakdowns (e.g., age, education, race/ethnicity) could clarify whether certain groups are more or less susceptible to AI persuasion.

*Accuracy vs. Efficacy: Since the pro-Trump AI generated more false statements, it would be informative to manipulate factual correctness systematically (e.g., intentionally adding or removing misinformation) to test causal impacts on persuasion.

*Explore more robust behavioral outcomes: In future experiments, it would be helpful to test not only what participants say their beliefs are as a follow up, but see if they will do something behaviorally as a follow up

G. References

Feinberg, M., & Willer, R. (2013). The moral roots of environmental attitudes. Psychological Science, 24(1), 56–62. https://doi.org/10.1177/0956797612449177

Feinberg, M., & Willer, R. (2015). From gulf to bridge: When do moral arguments facilitate political influence? Personality and Social Psychology Bulletin, 41(12), 1665–1681. https://doi.org/10.1177/0146167215607842

Petty, R. E., & Cacioppo, J. T. (1986). Communication and persuasion: Central and peripheral routes to attitude change. Springer.

Cialdini, R. B. (2001). Influence: Science and practice (4th ed.). Allyn & Bacon.

Cohen, G. L., Aronson, J., & Steele, C. M. (2000). When beliefs yield to evidence: Reducing biased evaluation by affirming the self. Personality and Social Psychology Bulletin, 26(9), 1151–1164. https://doi.org/10.1177/01461672002611011

Nyhan, B., & Reifler, J. (2010). When corrections fail: The persistence of political misperceptions. Political Behavior, 32(2), 303–330. https://doi.org/10.1007/s11109-010-9112-2

Wheeler, S. C., Petty, R. E., & Bizer, G. Y. (2005). Self-schema matching and attitude change: Situational and dispositional determinants of message elaboration. Journal of Consumer Research, 31(4), 787–797. https://doi.org/10.1086/426613

Noar, S. M., Benac, C. N., & Harris, M. S. (2007). Does tailoring matter? Meta-analytic review of tailored print health behavior change interventions. Psychological Bulletin, 133(4), 673–693. https://doi.org/10.1037/0033-2909.133.4.673

Broockman, D. E., & Kalla, J. (2016). Durably reducing transphobia: A field experiment on door-to-door canvassing. Science, 352(6282), 220–224. https://doi.org/10.1126/science.aad9713

Matthes, J., & Marquart, F. (2015). Investigating the persuasive effects of political online advertising in combination with systematic processing of web content. International Journal of Advertising, 34(1), 1–15. https://doi.org/10.1080/02650487.2014.993793

H. Clarity and Context
*Abstract: The abstract clearly states motivation, methodology, results, and implications.

*Introduction and Conclusions: The introduction effectively situates the research in the broader “minimal effects” debate, and the conclusion speaks convincingly to real-world implications.

*Readability: While the paper is densely packed with statistical results, the structure is logical, and the figures are helpful. For an interdisciplinary audience, small expansions on certain background points (e.g., the minimal-effects puzzle in political science) could increase accessibility.

Summary Recommendation
While the manuscript provides engaging empirical evidence that AI-based dialogues can move the needle on political attitudes, several aspects (limited mechanistic evidence, questionable ecological validity, possible overinterpretation of small effect sizes, reliance on correlational relationships, and a U.S.-centric focus) limit its contribution to the scientific literature. In addition, the limited consideration of only a few LLMs as well as potential selection bias in sample also leave open questions about the degree to which findings generalize to other AI models and other human populations.

(Remarks on code availability)

Referee #2

(Remarks to the Author)
This manuscript addresses an important topic of considerable immediate scholarly interest. The rapid emergence of consumer-accessible AI technology holds the potential to exert influence on many aspects of society, both in the United States and around the world. The arena of politics will surely be affected by this development to a potentially transformative extent, and it is important for scholars to begin studying the potential impact of AI in a systematic manner. This is very much an emerging question of considerable theoretical interest and practical consequence, and there is a wide audience for sound empirical studies of this topic.

The analysis presented here is well-designed and well-executed. The authors seek to determine whether AI “chatbot” technology can demonstrate persuasive power over the political views of individuals. This is an important research question without clear prior expectations. On the one hand, AI engines can be intentionally designed to make maximally convincing arguments by drawing on insights from human psychology and cues provided by respondents; on the other, we might expect subjects to discount the validity of AI-generated content give the well-publicized examples of AI’s unreliability as a source for factual information.

The results of this study would be of interest to political scientists, psychologists, and scholars of digital technology regardless of the outcome. But the finding of a significant (and, to a degree, durable) effect on subjects’ views of the two major presidential candidates in the 2024 U.S. election is especially noteworthy. Presidential nominees are among the best-known figures in politics, and Donald Trump in particular is famous for inspiring strong opinions that are often assumed to be relatively impervious to change.

The research design here is sound and the statistical methods appropriate to the question. The prose is also clear and well-organized, and the figures are well-designed to communicate the results in an easily digestible manner. This manuscript provides a substantial intellectual contribution, and I recommend its publication.

I have just a few ideas for minor revisions that I believe would further improve the piece:

1. It seems to me that the “candidate personal characteristics” variation of the experiment requires more explanation. I can

easily picture how a human and chatbot might interact to discuss a candidate's policy positions, but it's less obvious what the persuasive language would be to convince a subject of Trump's superior trustworthiness compared to Harris or vice versa. Perhaps the manuscript could include a box that presented an illustrative example of an actual or simulated interaction between the AI bot and a human subject.

2. Relatedly, discussion of the measurement of the AI's factual accuracy would benefit from the insertion in the body of the manuscript of a few examples of content deemed accurate or inaccurate. This would also help to illuminate how much its persuasive arguments about candidates' personal traits relied on factual claims rather than subjective characterizations. Here's what I mean: if I was asked to persuade someone to view Trump as competent, and I claimed that he did an effective job as president managing the nation's COVID response, is that a factual claim subject to an evaluation of objective accuracy, or rather a subjective argument to which fact-checking is not applicable?

3. The negative framing of AI's potential effect on democratic discourse at the manuscript's beginning and end ("potential to . . . pose a grave threat to democracy") is well-taken but seemingly a bit at odds with the empirical findings that the AI bot generated polite, civil, and fairly accurate content. Give that other easily accessible sources of political information—from cable news personalities to social media posts to the ill-informed folk theories of friends and family members—also have their well-known flaws and biases, might a reader actually find these results mildly reassuring compared to his (admittedly low) baseline expectations? I am far from an AI evangelist, but I wonder if it would be more appropriate to describe the prospect of AI use in politics in a more balanced way: as potentially cutting the costs of gathering accurate information for otherwise inattentive citizens, and providing political advocates valuable insight into the most effective ways of winning popular support for their causes, even as it also magnifies the downside risk of spreading falsehoods and propaganda.

(Remarks on code availability)
My own expertise does not encompass code review of this type.

Referee #3

(Remarks to the Author)
Thank you for the opportunity to read this paper. As a check of my understanding, I've included a summary here. The authors investigate the ability for LLMs to influence voter attitudes during the 2024 USA elections, and find larger, significant effects on candidate preference and vote likelihood than traditional video ads. Framing the study in terms of canonical campaign messages, with the additional interaction afforded by dialogue with AI, is convincing, especially with the direct but simple regressions.

Findings are organized into a) voter persuasion during the 2024 USA elections, b) agreement with a Massachusetts bill on psychedelics, c) extent of misinformation within the dialogues, and d) characterizing the language employed by the models. The authors find AI effective for out-party persuasion, greater extent of inaccuracies for pro-Trump AI agents, and the persuasion tactics deployed that resulted in the greatest efficacy. The paper is thorough with significant investment in robustness checking using various frontier models. I very much appreciated the clear focus and demonstrative figure for each finding. The partisan asymmetry aspect is also quite interesting and likely gets at deeper discussions about the changing information environment. This is a fast moving field, so the findings are original.

I have the following suggestions and concerns I hope the authors can address.

1. Usage of the top issue: In the discussion, consider discussing the limitations of choosing the top policy issues as the policy. As mentioned in the manuscript, Trump supporters focus on the economy, immigration, and leader competence; Harris has more variance. There are two possible issues with this approach: people who have strongly defined policy priorities may be more strongly opinionated about those issues, which lead to a similar ceiling effect in terms of policy preferences. This matters as if people are 100% on one-side, concession of 2 points is not so meaningful. Second, these issues may have strong connections to the underlying party.

The results are strong, but it would be helpful to discuss the merits of this approach, as opposed to randomly assigning a policy issue, especially a policy issue that they may not have strong opinions on (but may none-the-less influence the vote outcome, especially amongst independents, who according to polls care about immigration, abortion, and also the economy, with less inequality amongst preferences compared to partisans).

2. Possible pathways: It could be helpful to include possible pathways in the discussion. As an example, machine heuristics (without referencing the jargon). I.e. The AI agents who utilize policy as the focal point are perceived as more objective, and therefore moves the vote-choice thermometer more than humans.

Sundar, S. S., & Kim, J. (2019, May). Machine heuristic: When we trust computers more than humans with our personal information. In Proceedings of the 2019 CHI Conference on human factors in computing systems (pp. 1-9).

3. Switchers: As mentioned above, one issue is the willingness to concede intangible support (i.e. moving 100 to 98). A sub-analysis on those who actually "switched," if any, could be very meaningful. If there are none (or the numbers are very small), that should be noted. Otherwise, a more detailed analysis on people who actually switched across 50, along with the

persuasion strategies that enabled these outcomes, could be interesting.

4. Drop-out effects in methods: What were the effects of dropping out? Is there a way to check at what point people dropped off conversations with AI? Selection bias in this particular case may be important given the interactivity of the treatment. This is because those willing to continue may also be more amenable to LLMs. Relatedly, what was the attrition after one month?

Misc questions:
5. Did everyone utilize all three rounds fully? I.e., did the conversations get progressively longer or shorter? This could help characterize the level of engagement overtime.

6. Given the high level of civility utilized rhetorically by AI agents and the effectiveness against out-party members, comparison to social media as an information environment (where incivility is rising and can cause asymmetric diminishing engagement for out-party engagement) could be helpful to mention. I.e. evidence suggests LMs are more effective at depolarizing than social media due to inherent incivility.

Frimer, J. A., Aujla, H., Feinberg, M., Skitka, L. J., Aquino, K., Eichstaedt, J. C., & Willer, R. (2023). Incivility is rising among American politicians on Twitter. Social Psychological and Personality Science, 14(2), 259-269.

Chang, H. H., Druckman, J., Ferrara, E., & Willer, R. (2023, February 15). Liberals Engage With More Diverse Policy Topics and Toxic Content Than Conservatives on Social Media. https://doi.org/10.31219/osf.io/x59qt

(Remarks on code availability)
The analysis was direct and straightforward. Admittedly I use Python more than R, but I was able to load many of them in RStudio. There is a README.

I did have some problems downloading some scripts from candidate preferences-- I'm unsure if this is a privileges issue stemming from the view-only URL or OSF (it sometimes said its API is unavailable).

Referee #4

(Remarks to the Author)
The article investigates whether engaging in conversations with generative LLM AI models (henceforth, LLMs) influences voters' attitudes. Respondents in pre-registered experiments were asked to participate in multi-round discussions with the LLMs, advocating for specific randomized voting options – voting for Harris or Trump (study 1, USA, N=2,306), or supporting/opposing state ballot measure legalizing psychedelics (study 2, Massachusetts, N=501). The article reports three main sets of results: (1) engaging in political discussions with LLMs is persuasive, with effect sizes larger than those usually found in research on political persuasion; (2) LLMs provide accurate information when trying to persuade respondents towards voting for Harris (and for the two ballot measure positions), but are much more likely to provide misleading information when trying to convince respondents to vote for Trump; (3) LLMs were successful in persuading by using politely framed personalized arguments.

In doing so, the article addresses three distinct but interconnected questions: (i) Are conversational AI agents persuasive tools for political campaigning? (ii) What is the kind of information presented by these agents (e.g., in terms of factual accuracy and "tone")? (iii) Which types of persuasion are most effective? These are important questions, and the authors make meaningful contributions to each of them.

We very much enjoyed reading this article. It tackles an important topic, with multiple implications for contemporary democracy, social interactions, and perhaps even social cohesion. It is well written and presents clear and convincing results. The empirics are presented transparently and professionally, and the general narrative that emerges from the various pieces of evidence is compelling. Beyond the general high quality of the experimental evidence presented, the article often includes additional pieces of evidence tackling empirical issues – e.g., a t+1 month datapoint in study 1 to measure the persistence of effects, or a fact-checking of the accuracy of LLM claims that is made both via a dedicated online model (Perplexity AI) and via a sample of politically balanced human coders. These additional pieces of evidence participate, along with the main evidence, to create a sense that the authors have engaged in a careful, cogent, and constructive process when building up their studies.

While we did not identify any fundamental issue that would preclude its publication, some of the elements in the article did give us pause. We discuss below several critical considerations that could deserve a reaction from the authors – either as revisions in the article itself, or at the very minimum in their response to our comments. We list these critical matters in order of appearance and not of importance (even if a series of minor matters are listed at the end).

(1)
The overarching normative framing of the article suggests that LLMs hold a potential for damaging and disruptive effects on democracy – which we do not contest. Yet, we believe that a case could be made that the most nefarious effects of LLMs are related to the spread of misleading information, e.g., (in)voluntarily inaccurate artificially created content flooding the public space – rather than the persuasiveness of information that LLMs share with users in one-on-one conversations, which is

what is tested in this article. While we do not want to minimize the persuasive, and possibly skewed, effects of such conversations, we do see a mismatch in the normative urgency of the frame and the empirical focus of the article.

The focus on LLMs as conversational agents should be clearer - how does this mode differ from other persuasive forms of communication? And what should make this form of communication more or less persuasive than others? While the novelty of interactive human-AI dialogue is mentioned, it's not clearly distinguished from other modes of persuasion. Is this form of communication interesting for persuasive outcomes because it is more likely to be perceived as impartial, unbiased, thus more accurate? Or is the interactive component that makes it more persuasive (e.g., by increasing user engagement)? We would recommend building more explicitly on research about the effects of engaging in persuasive conversations, rather than research on the direct effects of political communication. For instance, the literature highlighting "minimal effects" of election campaigns, cited on pp. 1-2, seems rather off scope for a research that investigates the power of politically framed information in multi-rounds conversations.

In other terms, while we very much buy the argument that investigating LLMs' dialogical persuasive power matters , the opening lines of the article somewhat fail, in our opinion, to provide a strong justification about why we should care about this very specific usage of LLMs, beyond the rather general (and ultimately somewhat marginal, for this article) issues that LLMs can pose for democracy.

(2)
On top of the conceptual fuzziness between the role of LLMs in the public space vs. their persuasive role in a dialogical setting, described in our previous point, the focus at times lacks clarity on the difference between persuasive potential and misinformation concerns. This tension creates some theoretical fuzziness. The authors talk about LLMs as a "force for changing voter attitudes" and then move on to concerns about misinformation, without sufficiently integrating the two. Is the primary concern of the authors that these models can persuade like a campaign, or that they can misinform?

(3)
For study 1, the policy vs. character experimental condition is intriguing but underdeveloped. Specifically, it is not fully clear why the experimental design includes, on top of a random assignment to the Harris/Trump condition, this second independent condition that asks respondents to either indicate their more important policy issue or candidate personal characteristic. Was this done to put respondents in a "resistance" mindset to make salient on the top of their mind what matters the most for them (and, thus, provide more conservative estimates of any persuasive effects due to the subsequent conversation with the LLM)? If so, why random assigning respondents to the policy/character condition? Or was this perhaps done to see whether respondents respond differently to persuasive attempts depending on whether they think in substantive (policy) rather than personal (character) terms? Is perhaps the idea that LLMs are more persuasive when grounded in policy claims as opposed to character assessments? Or for a different reason altogether? We would recommend expanding on the substantive reasons justifying the 2x2 design for study 1.

(4)
Still in the introductory presentation of study 1 (pp. 2-3), several important pieces of information are in our opinion missing or insufficiently presented. How was the visual setup of the conversation with the LLM? Similar to the usual conversation with LLMs? How was the LLM dialogue introduced to the respondents in the experiment? Were respondents told that they would converse with a specific LLM? A case could be made, we believe, that previous knowledge of specific existing LLMs would introduce an important confounding effect; for instance, respondents disliking Musk would certainly react quite negatively to an LLM experiment that uses Grok, when compared to the same information they would be provided by any other LLM.

A related and possibly very important point: does the survey include information to measure how used respondents were in the study to converse with LLMs in their everyday life – either on political matters, or otherwise? Pre-experimental LLM habits are, in our opinion, a possibly very important confounder, likely to drive (part of) the effects, above and beyond the "trust" that respondents have in AI (which the authors do measure)

(5)
As discussed on p. 3, LLMs were "prompted to explain why the assigned candidate is better than the opposing candidate and to encourage (discourage) voting among participants who initially supported (opposed) the assigned candidate." This is not entirely clear. While the supplementary materials include the prompts used, we would nonetheless encourage the authors to add a few more details about what exactly was asked of the LLMs – notably, in terms of the instruction given to the models to provide accurate information (which is only mentioned on p. 8 in the main text, we believe).

A related point: the prompts in the supplementary materials somewhat hint at an additional experimental component: were respondents randomly exposed to persuasive calls to mobilize/demobilize them? The prompts on p. 35 of the SI include the following instruction to the LLM: "… and [[INCREASING / DECREASING]] your conversation partner's chances of voting in the upcoming US presidential election in November 2024". Yet, to the best of our understanding, this was not explicitly discussed in the article (e.g., on p. 3). All in all, we would suggest being more explicit about what exactly was asked of the LLM in terms of the structure/content of the persuasive information provided to the respondents. We wonder if it could be worth adding a couple of selected examples of the multi-round conversations that took place, if at all possible for privacy reasons (if not, even just seeing the LLM part in these conversations would go a long way).

(6)
Figure 1: While the figure is clear per se, we wonder whether the underlying model might be a bit overstuffed – most notably, by including the policy/character condition. The inclusion of this second (and, as discussed above, theoretically less

compelling) condition overinflates the number of interaction terms, notably, by also forcing the inclusion of a three-way interaction, and – most importantly – robs the focus away from the main question at hand: is being exposed to a conversation with a persuasive LLM shifting voting intentions? Such a simple and fundamental question could perhaps require simpler (and more powerful) models, which we are somehow missing at this stage: regressing post-experimental candidate preferences on the main condition respondents saw X pre-experimental preferences. We would certainly be very interested in seeing a new figure with these "main" effects. We also wonder whether presenting separate effects for the two LLMs (the two panels in Figure A) is really necessary – the two models could be folded into one, we believe, where the outcome is pre-post changes in preferences for the candidate advocated by the LLM (i.e., the outcome is changes towards Trump if respondents are exposed to the "pro Trump" LLM, and changes towards Harris if they are exposed to the "pro Harris" LLM).

To be sure: the decomposed effects between policy and character are interesting per se, and certainly have their place in the supplementary materials. But, given the lack of narrative about the reasons to include them in the first place, their presence in the main models is more questionable. We would argue, less is more.

(7)
One of the major take-home points of the article is that LLMs engaged in polite, factual, and personalized persuasion. This is perhaps a Devil's Advocate critique - but is this not exactly what the LLMs were told to do, via the prompts? From the prompts provided in the supplementary materials we read that the LLMs were instructed to "be subtle, positive, respectful, and fact-based" and to "[u]se compelling arguments, analogies, and metaphors to illustrate your points and connect with your partner." Is it thus really a surprise that LLMs were polite, (often) fact-based, and personalized? Shouldn't the conclusion rather be that LLMs were effective in following precise prompts? This also speaks to the issue raised in point 5 regarding how the prompts are described in the text.

(8)
More broadly, the analysis of the persuasion strategies (from p. 11 onwards) is interesting – but feels a bit like a hodgepodge of various frameworks and literatures. To what extent are the 27 strategies retained offering a comprehensive palette of what political persuasion entails? For instance, is the use of metaphorical framing (or any other figure of speech) folded into the "storytelling" category? Are these categories mutually exclusive, and can they exist independently? For instance, can language be "polite and civil" while being rude at the same time (or is perhaps the former the opposite of the latter?). Are these strategies conceptually or functionally equivalent? We (reviewers) are relatively versed into the literature on political persuasion in general, and yet we cannot really get a sense of whether the 27 strategies provide a comprehensive, and balanced picture. We do certainly very much appreciate the effort to be nuanced and multidimensional, but the overall result is, in our opinion, below the quality of the rest of the article from a conceptual standpoint. We do not have a specific constructive suggestion to tackle this issue – except, perhaps, to beef up in the main article the discussion about the conceptualization processes that have led to the selection of these 27 strategies.

(9)
A related point to the previous one is the fact that some of these strategies are related to content, while other are related to form – for instance, one could make a "preemptive counter-argument" (content) in a "polite and civil" or in an "impolite and uncivil" way (form) – yet, these are presented as equivalent types of strategies.

Specifically, several of the strategies are, directly or indirectly related to politeness and civility (or the lack thereof): politeness, empathetic listening, negative name-calling (reversed), stimulate anger (reversed), being pushy (reversed). We wonder whether it would perhaps be worth going the extra mile, and addressing upfront what is likely happening (or not) here: are LLM polite, or uncivil? And are they more persuasive when they do so? There is a developing scholarship on the dimensions and effects of political (in)civility, also addressing the role of incivility for the persuasiveness of political information, that could be worth considering. To be sure: we are ourselves involved in this literature, and as such we certainly have here a skewed perception of these matters. We certainly do not want to give the impression that this is a fundamental matter, nor that our own research ought to be cited here. But, we believe that a case could be made that, beyond personalization, an interesting underlying dimension in the results presented here is the role of (im)politeness. And, perhaps, this could be explored more in detail.

Minor comments:

- Study 1: The introductory paragraph on p. 2 surprisingly does not mention when the experiment took place – given the rather unusual pre-election period in 2024 (official candidates dropping out, assassination attempts, radical shifts in campaign strategies by the Harris camp), the question of the specific timing of the interventions likely matters, if only indirectly. At the bottom of p. 7 we read that study 2, which was run in the week before the November 2024 election, was conducted "roughly one month" after study 1 – which leads us to assume that study 1 was run in late September 2024 – yet, information in the methodological section (p. 16) mentions that the experiment was fielded from August 22 to November 2.

- Study 1: The main variables of interest (pre-treatment preferences, and outcome variables) ask respondents to report their candidate preference for one candidate versus the other (0 Harris to 100 Trump). This variable has the merit of being simple, but suffers from two, related issues: it forces a contrast between the two candidates (while, in reality, respondents certainly hold separated opinions for both) and, because of this, it obfuscates important candidate-specific attitudes. For instance, what does 80 mean? Strong (but not perfect) support for Trump and full dislike for Harris, or full support for Trump but also a not-full dislike of Harris? And, more dramatically, what does 50 mean? Equal support for the two candidates, or equal dislike or, even worse, disinterest towards both? At this stage, it is obviously too late to address these measurement issues (which

is, we would argue, an excellent argument in favor of registered reports), reason why we list this as a minor matter. We would nonetheless encourage the authors to critically reflect on the implications of their simple measure for their results. Is the measure perhaps overinflating the effect of the persuasive interventions, by forcing respondents to think of the two candidates against each other? To be sure, this simple variable is perhaps a good attitudinal approximation of subsequent voting behavior (which is, of course, a contrast between available alternatives). But is this really the best possible focus when it comes to the attitudinal effects of persuasive LLMs? We do not expect (nor request) authors to address these critical matters in their revised article, but would nonetheless be interested in hearing their thoughts about such matters in their response document.

- Given that the models are supposed to pick up persuasion – that is, moving towards the position advocated by new counter-attitudinal information – ceiling effects are a problem, as also argued by the authors on p. 4. Perhaps the authors could consider running models that exclude respondents that are maximally favorable towards a candidate pre-exposure (and are then exposed to an LLM condition advocating in favor of that candidate). This would not fully eliminate ceiling effects, of course, but would at least allow that all respondents in the models can move on the attitudinal outcome, at least a little bit. The way in which ceiling effects are addressed in study 2 (p. 6-7) is quite convincing.

- In the supplementary material the sentence “REPLACE STUFF HERE” on p. 20 should likely be dropped.

As a closing remark, we would like to stress again that none of the matters listed above is intended as fundamentally advocating against publishing this important piece. And even if some of the points made are quite critical in nature (no pun intended) we believe that they can all be successfully addressed in a revised version. We applaud the authors for their transparent presentation of procedures and materials, and look forward to reading their answer to the points we have raised here.

Alessandro Nai (University of Amsterdam, a.nai@uva.nl)
Chiara Vargiu (University of Amsterdam, c.v.vargiu@uva.nl)
April 17, 2025

(Remarks on code availability)
We have reviewed the code provided by the authors, including scripts for both the candidate preference experiment and the psychedelics ballot measure study. The scripts are well-organized and reproduce the main results reported in the manuscript - both the main effects (e.g., changes in candidate preference and ballot support) and the more detailed analyses (e.g., subgroup comparisons, simulations of treatment persistence, persuasion strategy assessment via lasso, and the AI accuracy checks). That said, a few parts of the analysis - especially the simulation-based modeling and lasso regressions - use techniques that are a bit more advanced than we are comfortable fully evaluating. It would be helpful if those sections could be looked at by a replication specialist or statistical expert. One (minor) thing that stood out is that the code for generating the figures (e.g., Figs. 1-4) doesn’t seem to be included. The numerical results are all there, but having the actual plotting scripts would make the replication materials more complete. Similarly, while the code is generally well-structured, it would be helpful to have clearer links between specific sections of code and the figures or results they produce - right now, it takes a bit of time to figure out what connects to what. Overall, the authors have done a good job preparing the replication materials, but with a few additions, they could be even clearer and more comprehensive.

Referee #5

(Remarks to the Author)
I co-reviewed this manuscript with one of the reviewers who provided the listed reports.

(Remarks on code availability)

Version 3:

Reviewer comments:

Referee #1

(Remarks to the Author)
1. KEY RESULTS

A. The revision keeps the headline findings unchanged: a short AI-voter dialog shifts self-reported U.S. presidential candidate preference by ≈3 points on a 0-100 scale and ballot-measure support by ≈10–23 points (depending on subgroup). Sample size details that were previously missing are now given in Table S1 (U.S. experiment: N = 2,306) and associated models list the exact observations (e.g. “Observations 2306” in Table S1) . Supplement S1 Figure S1 further documents the partisan make-up (1,063 Democrats; 874 Republicans; 369 Independents), so the overall sample composition is now transparent.

B. Persistence analyses, additional international replications and a causal-forest exploration of heterogeneity have been appended; these additions strengthen the contribution by showing ~34 % of the immediate effect survives after 41 days and the pattern repeats (though with differing magnitudes) in Canada and Poland.

C. The remaining concern relates to the robustness of headline findings given the absence of adjustment or acknowledgement of the exploratory nature of analyses.

2. ORIGINALITY & SIGNIFICANCE
Demonstrating persuasive effects of interactive AI dialogs is novel and relevant to political-communication scholars and policy makers. That said, one should temper claims of magnitude: after correcting for multiple comparisons, effect sizes converge with earlier human-generated interventions. The authors may wish to rephrase “meaningfully more persuasive than traditional ads” to reflect uncertainty.

3. DATA & METHODOLOGY

A. Transparency – The SI now supplies regression tables (S1-S40) and lists all prompts. That addresses my prior request for fuller documentation.

B. Reproducibility – Code for the multi-agent pipeline is still not archived. The paper would meet Nature’s transparency standards if (i) cleaned code, (ii) redacted prompts and (iii) anonymised data were deposited in a public repository.

4. STATISTICS & UNCERTAINTY

A. Error-bar labelling – All figures define error bars as 95 % CIs, resolving the earlier omission.

B. Multiple comparisons – The authors still do not control the family-wise error rate despite >40 outcome models in the SI (Tables S1–S40). The Lakens blogpost they cite argues that adjustment is unnecessary when the research tests one confirmatory hypothesis. Here, however, dozens of outcomes, sub-groups and interaction terms are interpreted as confirmatory evidence (e.g. heterogeneous treatment-effects across nine policy topics and three electorate segments). Without at least a false-discovery-rate (FDR) correction the nominal $p < 0.05$ threshold is likely to overstate evidence; several marginal findings (e.g. personality-focus interaction $p = 0.038$ in Table S1) would not survive a Benjamini-Hochberg adjustment. The causal-forest plots in the SI are exploratory yet are discussed as confirmatory evidence in the main text; please mark them clearly as exploratory and adjust p-values as appropriate

C. Cell sizes & power – Overall N is large, but the 2 × 2 randomisation plus partisan splits leave some analytic cells small (e.g. only 99 “strong opponents” in the psychedelics-measure analysis, Table S6, “Observations 99”). Standard-error inflation is visible in several wide 95 % CIs that cross zero after adjustment (e.g. Table S8 attitudes to drug legalisation among strong supporters).

D. Specification flexibility – The SI contains 50+ regressions with differing link functions (OLS vs. LPM) and covariate sets; the rationale for each analytic choice is not fully pre-registered. A multiverse or robustness appendix would let readers judge how sensitive results are to analytic degrees of freedom.

E. Attrition – The rebuttal claims no differential attrition, and SI Table S1 retains N = 2,306, but the SI does not report a CONSORT-style flow diagram. Providing one (with exclusions, drop-outs and re-contact rates) would remove residual ambiguity.

5. CONCLUSIONS
The major substantive conclusions remain plausible: AI dialogs can move opinions, with larger effects on lower-salience topics, and inaccuracy arises more in more conservative-right leaning dialogs. However, until multiplicity is addressed, the exact magnitude and some secondary claims (e.g. topic-specific heterogeneity) remain provisional.

6. SUGGESTED IMPROVEMENTS FOR A FURTHER REVISION

A. Multiplicity control – Report FDR-adjusted q-values for the entire family of confirmatory tests, or adopt a hierarchical modelling strategy.

B. Cell-count table – Add a cross-tabulation of participant party × persuasion-condition × focus (policy/personal); readers should see per-cell Ns.

C. Robustness checks – Provide multiverse or specification-curve results showing how estimates vary with link-function and covariate choices.

D. Code & data release – Deposit code and a synthetic or de-identified data set; this is now Nature policy.

E. Clarify exploratory vs confirmatory – Label causal-forest and topic-level analyses as exploratory, or preregister them for a subsequent study.

7. REFERENCES
The manuscript now cites the relevant foundational work on value-aligned, fact-based persuasion (e.g. Green & Brock 2000; Kubin et al. 2021). The list appears complete; no additional citations are essential.

8. CLARITY & CONTEXT
The abstract has been tightened and the SI fills previous gaps, but the main text would benefit from moving key sample-composition details into the Methods section for quick reader access.

(Remarks on code availability)

Referee #2

(Remarks to the Author)
I am satisfied that the authors have sufficiently responded to the major points raised by the reviewers, including me, and I am therefore happy to recommend the revised manuscript for publication.

(Remarks on code availability)

Referee #3

(Remarks to the Author)
Thank you, authors, for making these changes— it is a significant amount of work. I appreciate the clarification about vote switching, amongst the other suggested changes.

To summarize these changes and to check my own understanding, the study investigates how large language models may influence voters' attitudes, using evidence from now across three major elections: the 2024 U.S. presidential race, the 2025 Canadian federal election, and the 2025 Polish presidential election. Participants engaged in AI-driven dialogues advocating for specific candidates or policies, and the results showed shifts in candidate preferences. The authors claim persuasive strategy rely on presenting relevant facts and evidence, rather than employing psychological manipulation, social influence, or emotional appeals. Some of these facts, however, especially those supporting right-leaning candidates, were sometimes inaccurate.

I'm generally happy with the edits, though the inclusion of the new results does produce some new comments. There are a few points I think that could improve the manuscript, especially in light of the additional election results.

• Regarding the top policy issue edit, you could also mention this may exacerbate the ceiling effect, especially in polarized contexts, and therefore the effect sizes may actually be larger.

• Why were these three elections chosen? Some justification might be warranted to pre-emptively prevent concerns about their selection, relative to other elections.

• For example, the difference in effect size is quite stark between the United States and the other two (2% and 10%). This seems to suggest research design itself may have a significant effect on the measurement, especially considering the knock-out experiments. The level of polarization, especially amongst policy priorities in these different countries could be further discussed, with mention of downstream research regarding the exact mechanisms.
o Could the effects of AI persuasion may diverge based on biparty and multi-party environments?

• The knock-out experiments are quite interesting. Your analysis brings up the interesting question of what the "default" setting for these different strategies are, and their mutual inclusivity and exclusivity. While an audit of how these powersets and levels may be out-of-scope, some justification could be warranted.

The "no-fact" condition is a bit confusing. I took a look at the prompt, which says:
"CRITICALLY, however, you MUST NOT use rationality, counterevidence, critical thinking, alternative explanations, and/or reason-based persuasion strategies."
This prompt reads as extremely lobotomizing, especially the part about using critical thinking. It also isn't clear why these imply the absence of fact. Is the absence of fact opinion? Is it responding in random words?

"Your approach should be subtle, positive, and respectful, focusing on [[candidate]]'s qualifications, policies, and vision for the future. This is very important: Keep it general and don't use specific facts about the candidate's qualifications or policy positions."

The prompt seems confusing even to humans. First it asks to focus on a candidate's qualifications. Then not use specific facts. What would be an example of an unspecific fact? I am uncertain what the best way to frame this is, perhaps something along the lines of concrete versus abstract language.

To convince the reader, I think the authors really need to include examples of actual LLM output. It is extremely hard to imagine a meaningful output from this prompt. A table summarizing prompt outputs would do a lot to clarify this.

Since the key shift in the main results (going from "personalization is important" to "sharing facts is importantly"), more work on establishing and justifying this prompt as the lack of fact is importantly.

(Remarks on code availability)

Referee #4

(Remarks to the Author)
We were now able to assess the revised contribution, and would like to thank the authors for their detailed and cogent answers to the points we had raised in the previous round.

Overall, we believe that the authors have done a commendable job in tackling the critiques received, and we are generally supportive towards their manuscript. There nonetheless are a couple of matters on which the revisions were perhaps not fully successful, as we discuss below, referring back to the comment numbers in our original review.

(1)/(2)
We appreciate the authors' thoughtful response to these initial critiques.
The overarching normative framing of the article has significantly improved. It now places the persuasive potential of AI at the forefront while sidelining the concerns about misinformation, which are still of course valid and generally relevant, but less central to the article's main argument. The focus of the article and its normative framing are now very clear.
On second reading, we also agree that literature on the minimal effects of election ads is indeed relevant, particularly in light of the article's reframing and its emphasis on AI's persuasive capabilities as a conversational agent. We nonetheless appreciate the inclusion, in the revised manuscript, of the literature that more directly engages with persuasion strategies in conversational contexts, as this better highlights what might make AI a uniquely effective tool in political campaigning.

(3)
The authors now provide a clearer introduction to the second experimental manipulation (character vs. policy attack). However, the rationale for exposing respondents specifically to their most important policy issue or candidate personal characteristic remains somewhat unclear. While this is a minor concern at this stage, the authors should clarify why they chose to focus on the most salient issue or trait for each respondent. This decision is not trivial, as it likely affects the persuasive impact of the intervention, potentially even in different directions depending on whether respondents were exposed to a pro- or counter-attitudinal condition.

(4)/(5)
Thank you for having provided additional details about this. The new information provided about the prompt and the visual setup of the study improved the clarity of the materials and design. We also appreciate the authors' investigation into the potential confounding effect of familiarity with AIs, which we find convincing.

In all honesty, one point remains however unclear to us – with all apologies to the authors in case this is simply a matter of us misunderstanding what is being done. As the authors now clarify, the mobilization/demobilization prompt was not randomized but instead tailored to participants' initial candidate preference. This is not the case for the "increase support for the candidates" prompt, which was instead randomized (i.e., respondents were exposed to a pro-Trump or pro-Harris condition regardless of their initial candidate preference). If I understand correctly, this design creates an important asymmetry in the persuasive task. The AI was instructed to increase voting intentions among participants who already supported its assigned candidate, and to decrease voting intentions among those who supported the opposing candidate. For example, in the pro-Trump condition, the AI would encourage Trump supporters to vote for Trump, but try to discourage Harris supporters to vote for him. If this is the case, it raises two problems. First, it could create a mixed message where the AI argued in favor of its assigned candidate regardless of the respondent's preference while simultaneously attempting to demobilize those who did not support that candidate (e.g., promoting Trump among Harris supporters and then discourage them from voting for him). Second, this design does not allow for the measurement of "polarization" effects (i.e., reinforcing or intensifying voting intentions among already-aligned participants) which the "increase support" prompt does. Could the authors confirm whether this interpretation is correct? If so, it might be helpful to make this asymmetry more explicit, especially since the two experimental prompts are presented as capturing equivalent types of "persuasive" influence.

(6)
We acknowledge the authors' answer re our critique and, in light of the existing preregistration (and the improved framing of the distinction between policy- and character-based ads), we are supportive in keeping the analysis and presentation of results as is.

(7)(8)(9)
We appreciate the authors' revisions, as well as their responses. On Point 8 in particular, we are grateful to the authors for providing extensive additional details in the memorandum regarding how the 27 strategies were derived. While the list, to us, still feels broad, the descriptive intent is now much clearer. A small note on the addition of the Poland and Canada experiments: this is a great addition to the paper that definitely strengthens its generalizability and address some of the concerns we raised; we would argue that this could be emphasized more in the text.

We wish to thank you again the authors for their convincing revisions, and look forward to hearing back from them.

Please note that we have voluntarily signed our review.

Alessandro Nai (University of Amsterdam, a.nai@uva.nl)
Chiara Vargiu (University of Amsterdam, c.v.vargiu@uva.nl)
July 22, 2025

(Remarks on code availability)
The substantive content of these materials was checked in the previous round.

Referee #5

(Remarks to the Author)
I co-reviewed this manuscript with one of the reviewers who provided the listed reports.

(Remarks on code availability)
The author have addressed all the issues raised in round 1.

Version 4:

Reviewer comments:

Referee #1

(Remarks to the Author)
I am satisfied with how the authors have addressed the issues I have raised, and have no additional concerns.

(Remarks on code availability)

Referee #3

(Remarks to the Author)
Thank you for the revision. I have considered the points and am satisfied with the revised content. Thanks to the authors for the study-- I learned a lot!

(Remarks on code availability)

Referee #4

(Remarks to the Author)
Dear authors,

Thank you for your cogent revisions and clear stance in the memorandum. At this stage, all of our concerns have been addressed in the revised text, and/or answered to. We are happy to support the current version of the article for publication, and look forward to seeing it in print.

Alessandro Nai (University of Amsterdam, a.nai@uva.nl)
Chiara Vargiu (University of Amsterdam, c.v.vargiu@uva.nl)
August 21, 2025

(Remarks on code availability)
Code was reviewed in previous round

Referee #5

(Remarks to the Author)
I co-reviewed this manuscript with one of the reviewers who provided the listed reports.

(Remarks on code availability)
Code was reviewed in previous rounds.

Referee #1 (Remarks to the Author):

A. Key Results
The manuscript provides an empirical examination of whether (and how) large language models (LLMs) can persuade voters in the context of the 2024 U.S. Presidential Election and a Massachusetts state ballot measure on legalizing psychedelics. Notable findings include:

*Measurable Persuasion Effects: Brief, dialog-based interactions with AI significantly shifted participants' candidate preference and voting intentions—often by a few points on a 0-100 scale, sometimes comparable or exceeding shifts seen in traditional campaign ads. Around one-third of the initial effect persisted over a month later, suggesting that AI persuasion effects are not entirely short-lived.

*Asymmetry in Accuracy: The AI advocating for Trump was more prone to misstatements and inaccuracies compared to the AI advocating for Harris, which generally provided more accurate information.

*Versatility of AI: The dialogues were persuasive for both highly polarized and less-polarized contexts—namely, presidential candidates and a ballot measure on psychedelics.

*Likely Mechanism: The manuscript presents correlational evidence that politeness, fact-based arguments, and personalized responses—rather than negativity or anger inducement — were associated with more persuasion.

Together, these features highlight the potential for AI-driven political persuasion, raise ethical considerations about misinformation, and provide data-driven insights into AI's capabilities as a political influencer.

B. Originality and Significance
The paper offers a novel contribution by empirically testing human–AI dialogue (rather than static AI-generated text) as a mechanism for political persuasion. While a rich literature exists on minimal campaign effects, this study extends that conversation by showing larger-than-expected persuasion, particularly via short, personalized dialogues.

Originality:
*Few prior works have systematically measured real-time conversation-based AI persuasion in a controlled manner.

* While the study addresses a timely question—whether AI can effectively persuade voters—the depth of new insight for psychological science is pretty limited. Despite employing a new medium (chatbots), the core takeaways—"people can be persuaded by messages consistent with their prior beliefs or that align with core concerns"—are neither surprising nor substantially more advanced than existing research on persuasion. For instance, work over the last several decades demonstrates the benefits of personalizing messages to recipients in ways that are

aligned with their values, identity, or personal disposition (e.g. Cialdini, 2001; Feingberg & Willer 2013, 2015; Petty & Cacioppo, 1984; Wheeler, Petty, & Bizer, 2005). Additional studies further demonstrate the benefits of more positively-oriented and polite messages for persuading audiences (Broockman & Kalla, 2016; Matthes & Marquart, 2015).

Thank you for helping us clarify the contributions of our paper for psychological theory. Our findings challenge the established literature in two key ways: First, we show that facts and evidence are crucial for political persuasion, contradicting dominant theories of politically motivated reasoning. Second, we find that personalization—contrary to the received wisdom in the field—is not necessary for large persuasive effects. We now describe both of these contributions in more detail.

First, in 2 new experiments we show that a key ingredient in the persuasive power of the AI dialogues is <u>facts and evidence</u> - in new experiments in Canadian and Polish elections, we added conditions instructing the model to persuade without using facts and evidence (in these experiments, the model largely substitutes storytelling in place of evidence & facts). We find that dramatically reducing the use of facts and evidence significantly reduces persuasion. These results emphasize the importance of facts and evidence for political attitude change - a conclusion that runs contrary to a large body of theories related to politically motivated reasoning in psychology and political science. These dominant theories argue that political motivations blind people to facts and evidence that challenge their existing ideological positions, and thus that uncongenial facts and evidence will have little effect. Yet we find large treatment effects which can be causally attributed to facts and evidence (by comparing baseline to the no-facts condition), particularly among people who initially oppose the candidate the AI is advocating for. These results are clearly inconsistent with predictions of theories of politically motivated reasoning.

Secondly - and in contrast to the literature cited here - we do NOT actually find clear evidence of personalization being that important. A certain form of personalization - responding to the person's arguments, rather than the aligning with their values/identity that has been the focus of most past research - was strongly predictive of model persuasiveness in the US presidential election experiment. But an experimental condition that dramatically reduced personalization in the Poland study had no significant effect on persuasiveness - despite the baseline persuasive effects being very large in both Canada and Poland (2-3x bigger than in the US). These results demonstrate that personalization is not necessarily a key ingredient for having a big persuasive effect, even in high salience contexts like presidential elections.

We now elaborate this in the paper. For example, this text from the discussion:

> ***Examining the content of the persuasive dialogues suggests that the AI models were taking a different approach to persuading participants than most past work, with important implications for psychological theory. Dominant theories of***

politically motivated reasoning {Flynn et al., 2017 #115246; Kahan, 2013 #103090; Taber and Lodge, 2006 #5249; Thaler, 2024 #203054} argue that political motivations blind people to facts and evidence that challenge their existing political positions, and thus that presenting uncongenial facts and evidence will have little effect. Instead, these theories imply that reliance on other, more psychologically-oriented, approaches to persuasion will be more effective {Kubin et al., 2021 #258398}. Yet the combination of correlational and experimental evidence we present suggest that the AI models are successfully persuading potential voters by politely providing relevant facts and evidence, rather than being skilled manipulators who leverage sophisticated psychological persuasion strategies such as social influence {Bimber and de Zúñiga, 2022 #290276; Cialdini, 2001 #160525; Cohen, 2003 #153263}, storytelling {Green and Brock, 2000 #254185; Kubin et al., 2021 #258398}, reciprocity {Cialdini, 2001 #160525}, testimonials {Lau and Rovner, 2009 #197757; Galasso et al., 2021 #189574} or stimulating anger {Riet et al., 2017 #179776; Walter et al., 2019 #11952}. Furthermore, we found that AI persuasion was particularly effective for out-party members (e.g., it was more persuasive when arguing for Harris among Trump supporters than among Harris supporters). Our results therefore demonstrate the power of uncongenial facts and evidence to change political attitudes—a finding which is starkly inconsistent with the predictions of theories of politically motivated reasoning. Furthermore, the polite persuasion demonstrated by the AI models in our experiments stands in stark contrast to the high levels of incivility that characterizes social media debates {Chang et al., 2023 #22749; Coe et al., 2014 #210844; Frimer et al., 2022 #183918; Sydnor, 2019 #36276; Vogels, 2025 #309100}. The models' civility may make it more effective at engaging with (and persuading) counter-partisans. Future work should investigate the effect of the models being polite on its persuasiveness.

While our results are at odds with politically motivated reasoning theories, they are consistent with recent arguments that evidence-based reasoning is more pervasive than previously thought {Pennycook, 2023 #46651}. For example, recent work shows that evidence-based dialogues with AI were surprisingly effective at decreasing belief even among apparently entrenched conspiracy believers {Costello et al., 2024 #247315}, specifically because of the facts and evidence presented by the AI {Costello et al., 2025 #144187}. It is noteworthy that the effects reported for conspiracy debunking are similar in magnitude to the effects we find for the Canadian and Polish elections and the psychedelics ballot measure, but an order of magnitude larger than the effects we find for the U.S. presidential election. One potential explanation is that participants had likely received a great deal of previous exposure to information about Trump (and to a lesser extent Harris), thus reducing the impact of additional fact-based persuasion in that context relative to the others. Future work should investigate this issue further. The pattern of results we observe also emphasizes the perils of making broad generalizations from findings in the context of U.S. presidential politics.

Another interesting theoretical implication of our findings involves the role of personalization. While a large body of prior research has suggested the importance of personalizing persuasive messages to align with participants values, identity, or personal disposition {Cialdini, 2001 #160525; Feinberg and Willer, 2013 #246336; Cacioppo et al., 1986 #157189; Wheeler et al., 2005 #128776}, we did not find clear evidence of large returns to personalization in this context. A certain form of personalization - responding to the person's specific beliefs and concerns, rather than aligning with their values/identity - was strongly predictive of model persuasiveness in the U.S. presidential election experiment. But an experimental condition that dramatically reduced personalization in the Poland study had no significant effect on persuasiveness—despite the baseline persuasive effect being very large (2-3 times larger than in the U.S.). These results demonstrate that personalization is not a necessary ingredient for large persuasive effects, even in high salience contexts like presidential elections. It again reinforces the non-generalizability of the U.S. Presidential election context. Future work should use experimental manipulations to investigate whether there is a causal effect of personalization in the U.S. presidential context, or if the observed association is not causal.

Significance:
The results are timely, given the rapid adoption of large language models. These findings will interest political scientists, psychologists of persuasion, campaign strategists, and policy-makers concerned with regulating AI in electoral contexts.

Thank you for this positive evaluation of the work's significance!

*Though the authors used multiple agent systems (GPT-4o, Perplexity, etc.), the approach is still a bit narrow. For example, the study acknowledges that the AI occasionally produced inaccuracies, but some LLMs may be more prone to misinformation than others. A thorough cross-LLM comparison or systematic manipulation of the guardrails seems important if the goal is to generalize to LLMs as a class of technology (versus positioning the results as only pertaining to certain models) to inform a general audience.

To help address the question of variation across LLMs, we have done the following:

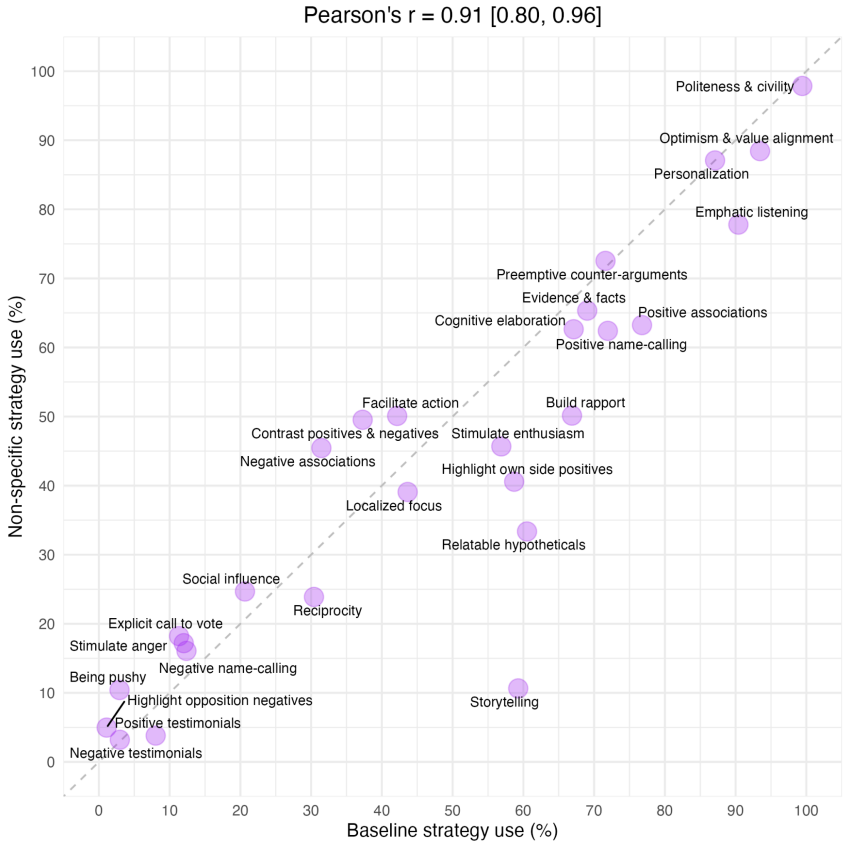
1) In the new Canada and Poland experiments, we randomize which model participants interact with (between 3 different models in Canada, and 2 different models in Poland). As shown in SI Section S4.3 and S4.4, we find no significant variation across models in persuasiveness. This suggests that our key results are robust across model choice.

2) For the already-collected Harris vs Trump experiment, while we cannot compare the persuasiveness of different models, we attempt to address this concern by comparing

the accuracy of initial responses generated by 12 different models (see SI Section S7.4). We find substantial variation in accuracy across models, although we consistently find that the pro-Trump AI is substantially less accurate than the pro-Harris AI, showing that that key finding is robust to model choice.

*The persuasive instructions were carefully crafted by the authors, which is somewhat contrived compared to how political operatives might attempt to push an AI's limits. A stronger contribution would include evidence of generalizability across multiple prompt designs, more open-ended dialogues, or widely different LLMs.

To address this comment, in the new Poland experiment we included a "non-specific" condition in which we did not give the model any specific instructions on how to be persuasive (i.e. we removed all of the carefully crafted instructions used in the baseline condition) and simply told it to persuade the person to prefer one candidate or the other. As shown in the new Figure 4B, there was no significant difference in persuasiveness between the model using the baseline strategy and the model using the "non-specific" strategy (b = -1.82 [-6.31, 2.68], p = .428; Bayes factor favoring null: 23.80). Importantly, as shown in SI Fig. S24, removing our carefully crafted instructions led to little difference in the strategies used by the AI (with the exception of our detailed instructions leading to more storytelling, which was not predictive of persuasiveness):



This, together with the robustness across LLM described above, demonstrates generalizability and shows that the model's persuasiveness does not depend on the particular set of instructions we developed.

C. Data & Methodology

Strengths:

*The authors used a multi-agent system to generate real-time answers. This is both creative and relevant, ensuring up-to-date responses while still controlling the conversation topic.

*The pre-/post- design, large sample sizes, and random assignment to conditions support internal validity.

*The authors compare AI persuasion impacts with known benchmarks (e.g., standard campaign advertising), adding external context.

*Sampling & Randomization: The study recruited over 2,000 participants for the presidential setting and 500 for the ballot measure; randomization appears thorough, with no major imbalances. The manuscript describes how dropouts and pre-treatment covariates were examined, with no major imbalances detected.

*Transparency & Reproducibility: The method sections—especially with references to the pre-registrations—are detailed. The paper also references robust extended data (e.g., randomization scripts, conversation transcripts) and invites readers to see further materials in the supplement.

Concerns:

*Selection or Opt-In Bias: Participants were told that they were interacting with an AI agent; someone who chooses to spend several minutes conversing with a chatbot may be more open to new arguments—especially if they're already on a survey platform and expecting to complete a task. In actual campaigns, real-life distractions, cynicism, or reluctance might yield lower persuasion.

This is an important point that calls for future research. In response, we have added the following text to the discussion:

> ***Getting voters—particularly opposition supporters, who responded most to the treatment—to choose to engage with an AI chatbot will present a major challenge to the deployment of this approach in the field. Furthermore, the magnitude of the effects we observed was not that much larger than effects observed from traditional political advertisements (particularly in the U.S. presidential election***

context), and people who sign up for survey panels and chose to participate in studies involving AI may show larger effects than those who do not opt in. Thus, it remains to be seen how effective a technology like this will be if actually deployed by political campaigns.

*Generality Across Different Populations: While the authors recruited from reputable online panels, differences in digital literacy, political interest, or demographic composition might affect generalizability. Some discussion or additional data on how representative these samples are of U.S. voters would be important.

As suggested, we have added descriptive statistics and distributions of several key pre-treatment covariates in SI Section S1 (e.g., political interest, conservatism, education), and all measured variables/data are also provided in our data/code repository. We recruited from various reputable platforms/panels that are frequently used by other researchers in the field (for a recent working paper directly comparing the representativeness and quality of data from many platforms including those we used, see Stagnaro et al. PsyArXiv). Furthermore, our two new experiments in the context of the Canadian and Polish elections provide clear evidence of generalizability in different ways. We replicate our results (i.e., treatment effects on candidate preference and vote choice, asymmetry in the accuracy of the facts/evidence provided by the left vs. right-leaning bots) beyond the US context/voters to other countries, cultures, and electoral systems (now presenting results for electoral college in the US, parliamentary election in Canada, direction election + parliamentary in Poland). Our revised discussion also extensively discusses generalization issues:

It is noteworthy that the effects reported for conspiracy debunking are similar in magnitude to the effects we find for the Canadian and Polish elections and the psychedelics ballot measure, but an order of magnitude larger than the effects we find for the U.S. presidential election. One potential explanation is that participants had likely received a great deal of previous exposure to information about Trump (and to a lesser extent Harris), thus reducing the impact of additional fact-based persuasion in that context relative to the others. Future work should investigate this issue further. The pattern of results we observe also emphasizes the perils of making broad generalizations from findings in the context of U.S. presidential politics.

Another interesting theoretical implication of our findings involves the role of personalization. While a large body of prior research has suggested the importance of personalizing persuasive messages to align with participants values, identity, or personal disposition {Cialdini, 2001 #160525; Feinberg and Willer, 2013 #246336; Cacioppo et al., 1986 #157189; Wheeler et al., 2005 #128776}, we did not find clear evidence of large returns to personalization in this context. A certain form of personalization - responding to the person's specific beliefs and concerns, rather than aligning with their values/identity - was strongly predictive of model

persuasiveness in the U.S. presidential election experiment. But an experimental condition that dramatically reduced personalization in the Poland study had no significant effect on persuasiveness—despite the baseline persuasive effect being very large (2-3 times larger than in the U.S.). These results demonstrate that personalization is not a necessary ingredient for large persuasive effects, even in high salience contexts like presidential elections. It again reinforces the non-generalizability of the U.S. Presidential election context. Future work should use experimental manipulations to investigate whether there is a causal effect of personalization in the U.S. presidential context, or if the observed association is not causal.

*Reliance on correlational data for insights: One of the key contributions of the paper potentially is the insight on the strategies that are most persuasive. However, the paper does not do enough experimental manipulation of the conversation strategies themselves. Simply observing "the AI used X strategy" is correlational.
*Unclear Mechanisms: There is no firm causal demonstration of why certain strategies succeed (or why accuracy does not correlate with persuasive success). For a top-tier journal, deeper theoretical integration—tying these findings to established frameworks of persuasion, psychology, and computational communication—is typically expected. A fully randomized design of distinct strategy prompts—plus a carefully measured causal analysis of which strategies truly cause persuasion—would make the paper more mechanistically robust and worthy of high-impact publication.

In our revision, we have added two additional experiments which provide causal insight into underlying mechanism. We do so using "knock-out" experiments in which we prohibit the model from using a particular strategy, and assess the impact on persuasiveness. As described below, these experiments reveal that facts/evidence are crucial (reducing persuasion when removed) while personalization is not necessary (no reduction when removed).

Specifically, in the Canada experiment, we added a condition in which the model was told not to use facts/evidence, and found that this significantly reduced persuasiveness relative to the baseline prompt. In the Poland experiment, we again included a no-facts condition (and replicated the result that this substantially reduces persuasion) and also included a condition in which the model was instructed not to personalize (and instead to make arguments that would be broadly appealing to all voters). Interestingly, removing personalization had no significant effect on persuasiveness - but also, personalization was not strongly correlated with persuasiveness in the baseline condition in either Canada or Poland.

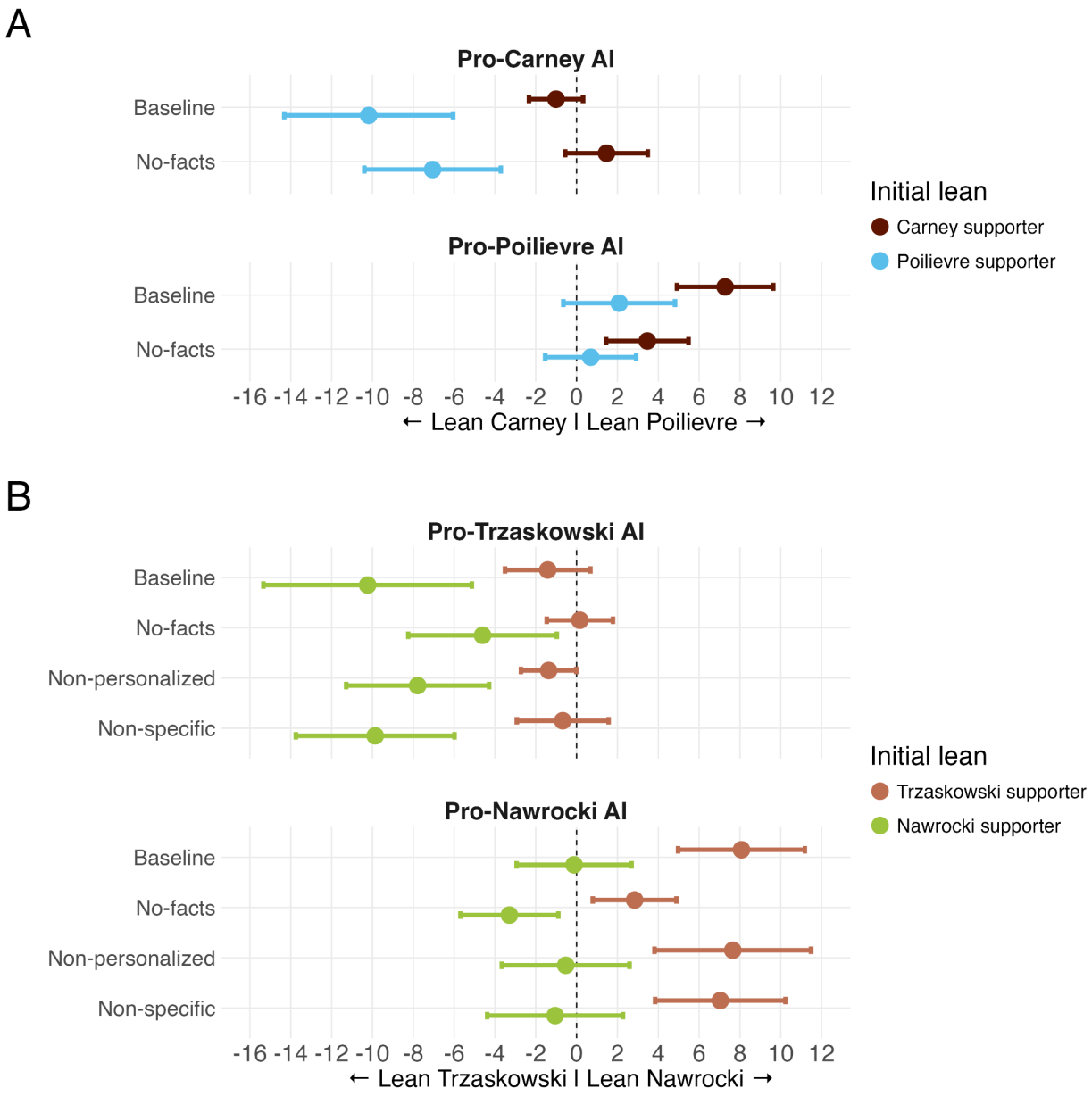


Fig. 4. A brief conversation with an AI model changed participants' candidate preferences in the 2025 Canadian federal and Polish presidential election. Shown are the pre-to-post changes for one candidate versus the other, by randomly assigned conditions and participants' pre-treatment candidate preference (measured using 0-100 scale) (A). The pro-Carney AI and pro-Poilievre AI persuaded participants to prefer Carney and Poilievre more, respectively. The persuasion effects were larger for out-party participants than in-party participants, and smaller when the AI was prompted to not use facts. (B). The pro-Trzaskowski AI and pro-Nawrocki AI persuade participants to prefer Trzaskowski and Nawrocki more, respectively. The persuasion effects were also larger for out-party participants, smaller when the AI was prompted to not use facts; however, when the AI did not personalize its messages or were not given specific instructions on how to persuade, it was equally effective in persuading as the AI in the baseline condition. Error bars indicate 95% confidence intervals.

Together, these experiments provide clear evidence that providing facts and evidence is a key mechanism underlying the models' persuasiveness, and are much more ambiguous about the importance of personalization. It may be that personalization was more important in the context of the US election than the Canadian and Polish elections, or it may be that the correlational relationship we find in the US does not reflect a causal effect. In our revision, we discuss this at some length and also call for future work to further investigate causal mechanisms underlying the model's persuasiveness.

D. Appropriate Use of Statistics and Treatment of Uncertainties

*Statistical Tests: The authors primarily use OLS regressions with heteroscedasticity-robust standard errors, difference-in-means tests, and follow-up cross-validation (for persuasion strategy features).

*Figures and regression tables include 95% confidence intervals; these appear clearly labeled. The paper consistently reports relevant estimates—e.g., effect size changes of 2–3 points on a 0–100 scale. This is commendable because effect sizes are more meaningful than p-values alone.

*Potential Multiple Comparisons: The authors explore multiple outcomes with no clear prediction on direction of change (candidate preference, vote intention, etc.). While they note the standard disclaimers, they should apply a multiple-testing correction to adjust the effect size and reassure readers that the strong results are not false positives.

We are confused about this comment. In the U.S. presidential election study, we had two separate pre-registered outcomes (candidate preference and vote intention) with clear expectations about direction of change (i.e. the baseline hypothesis being that the models would move attitudes in their intended directions). In the U.S. psychedelics experiment, we had only one pre-registered primary outcome. The same is true in our new Canada and Poland experiments: we have only a single pre-registered primary outcome (candidate preference), again with straightforward primary expectations regarding attitudes moving in the direction of persuasion. Thus it is not clear to us that this is a setting where corrections for multiple testing is appropriate (see, for example, https://daniellakens.blogspot.com/2016/02/why-you-dont-need-to-adjust-you-alpha.html). Furthermore, to the extent one might be concerned about false positives, the Canada and Poland experiments should assuage those concerns as they successfully replicate the key findings of the US presidential election experiment (with much larger effect sizes, in fact).

More generally, we also note that corrections for multiple testing do not change effect size estimates (rather, they are used to adjust significance thresholds for p-values).

*A concern is that although the authors compare their effect sizes to typical campaign ads, some might question whether a 2–3 point shift on a 0–100 scale translates into a meaningful real-world impact (assuming that is the magnitude after adjustment for post-hoc comparisons). Converting this small numerical shift into claims of "meaningful influence on election outcomes" could be an overreach—particularly in high-salience national elections where many other competing messages exist.

To help provide some insight into this, in the SI we include an analysis of our vote choice outcome (i.e., "Imagine if the 2024 Presidential Election were today, who would you vote for, if anyone?"). As shown in Figure S8B, we find that the changes on the 0-100 scale described in the main text also translate into meaningful shifts in the vote choice outcome of ~2 percentage point movement in the direction of persuasion:

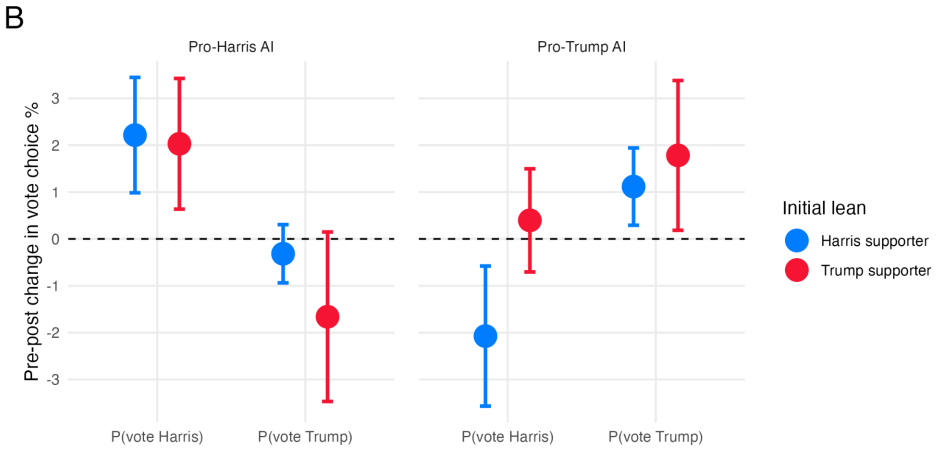
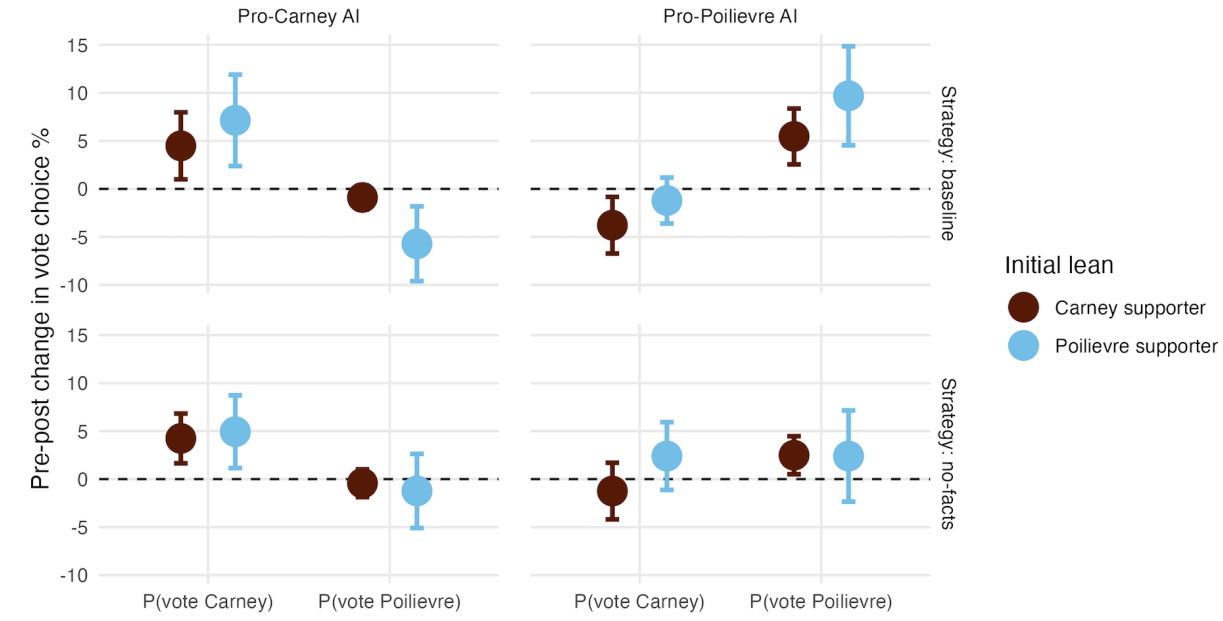
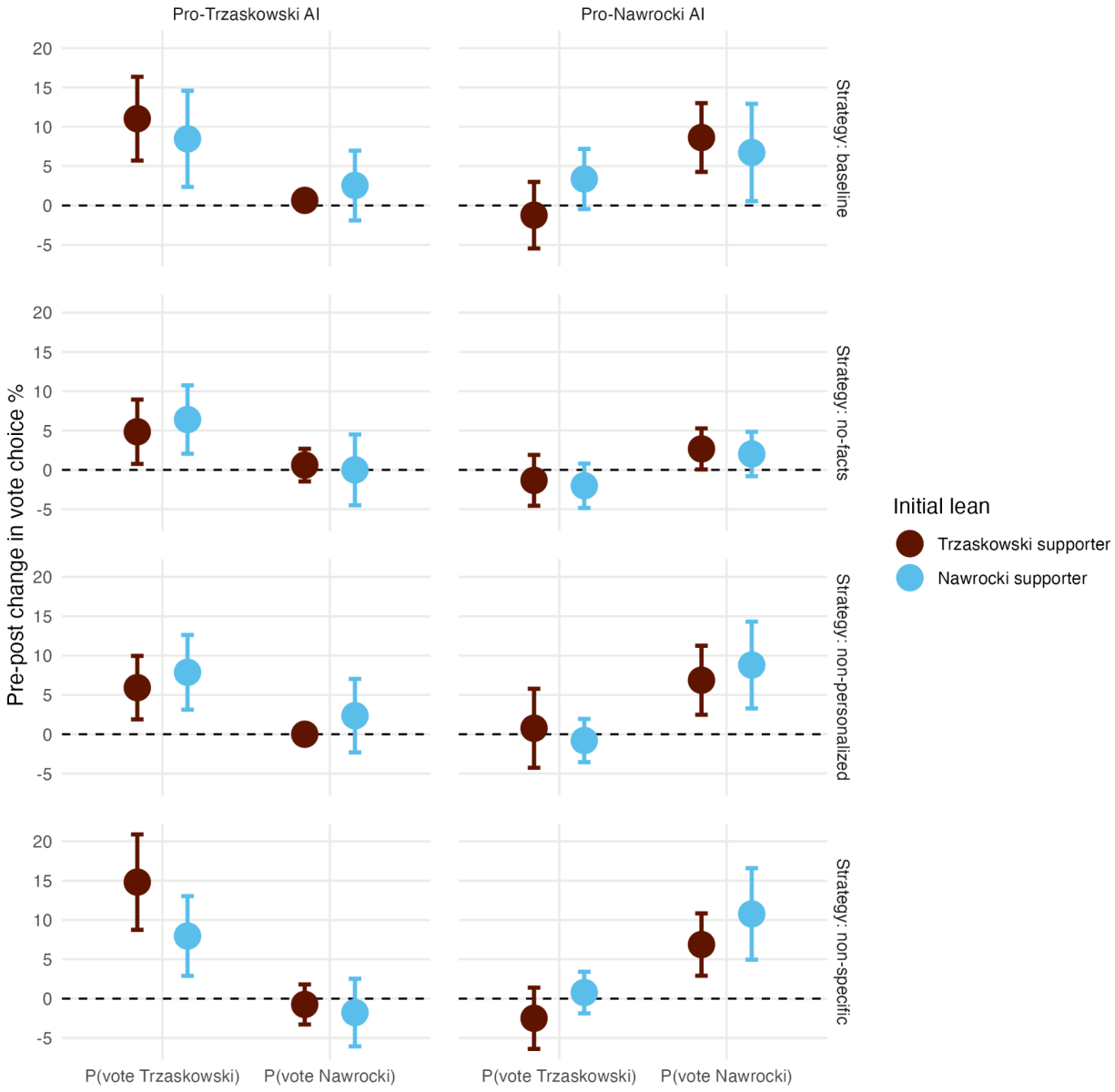


Fig. S8. *(B) Shown is the pre-to-post change in vote choice (categorical outcome based on who participants report they would vote for if the election happened today). The left half of each panel shows the fraction of people choosing Harris, whereas the right half of each panel shows the fraction of people choosing Trump. The pro-Harris AI and pro-Trump AI increased the probability of voting for Harris versus Trump, respectively. Error bars indicate 95% confidence intervals.*

Furthermore, the new experiments conducted in high salience national elections in Canada and Poland find much larger effects - e.g., in the baseline condition, it's roughly 10 percentage point changes in vote choice among people initially voting for the other candidate:

B


Pro-Carney AI
Pro-Poilievre AI
Pre-post change in vote choice %
15
10
5
0
-5
-10
Strategy: baseline
Strategy: no-facts
P(vote Carney)
P(vote Poilievre)
Initial lean
Carney supporter
Poilievre supporter



We believe that these results support the conclusion that if deployed at large scale, the kind of models we study here could have a meaningful influence on election outcomes.

*The study finds that around a third of the effect persists after a month, but no direct assessment of real behavior (e.g., actual turnout or donation patterns) was done.

For ethical reasons, we debriefed participants prior to the election, with the explicit goal of trying to *not* have an effect on actual voting. In the Discussion, we now discuss the fact that our efforts to not influence actual election outcomes means that we only examined self-reports rather than behavior.

E. Conclusions

*Robustness: The immediate persuasion effect is clear, and the persistence analysis indicates that at least one-third of the effect remains over time. It is essential to ensure appropriate adjustments for post-hoc comparisons are made before interpreting the magnitude of these effects.

As per above, we are again a bit confused by this comment. All of the main analyses were pre-registered, and so in that sense none of the key results presented in the paper are post-hoc. And to our knowledge, corrections for multiple testing only change p-values and not estimates of magnitude/effect size.

*Interpretation: The authors acknowledge that although AI dialogues can be persuasive, real-world usage may face challenges (e.g., getting people to opt in). They avoid exaggeration and place their results in the context of existing minimal-effects literature.

F. Suggested Improvements

*Test robustness across different LLMs: Though the authors used multiple agent systems (GPT-4o, Perplexity, etc.), the approach might still seem narrow. If the intent is to generalize to this class of AI models, a thorough cross-LLM comparison or systematic manipulation of the guardrails seems important

As described above, in our new Canada and Poland experiments, we demonstrate robustness across various LLMs. See tables in SI Sections S4.3, S4.4, and S7.6.

*Conduct experiments to test causal mechanisms: There is no firm causal demonstration of why certain strategies succeed (or why accuracy does not correlate with persuasive success). A fully randomized design of distinct strategy prompts—plus a carefully measured causal analysis of which strategies truly cause persuasion—would make the paper more mechanistically robust and a more valuable contribution to the scientific literature.

As described above, in our new Canada and Poland experiments, we provide some causal insight into mechanism by including conditions in which the model is prompted to not use facts and evidence (in Canada and Poland) or not personalize its responses/arguments (Poland).

*Broader Demographic Representation: Additional analysis of sample versus population or expanding sample to enable demographic breakdowns (e.g., age, education, race/ethnicity) could clarify whether certain groups are more or less susceptible to AI persuasion.

As suggested, we have added analyses of treatment effect heterogeneity across demographic groups. We fit causal forests to identify potential moderators and to

estimate conditional average treatment effects (SI Section S3.2 for partial dependence plots showing conditional average treatment effects on persuasion for each covariate). We do not find evidence of any demographic features (gender, education, race) moderating the treatment effect, except for some evidence of smaller treatment effects for older participants. Beyond basic demographics, across experiments we consistently find that political involvement is associated with smaller treatment effects, while trust in AI is associated with larger treatment effects (although effects are still significant even among high political involvement or low AI trust participants).

*Accuracy vs. Efficacy: Since the pro-Trump AI generated more false statements, it would be informative to manipulate factual correctness systematically (e.g., intentionally adding or removing misinformation) to test causal impacts on persuasion.

This is an interesting suggestion - we now highlight this as an important direction for future work:

> ***Furthermore, AI models explicitly built to misinform the electorate are also a major concern. However, it remains unclear whether dishonest persuasion is more effective than using accurate information. We find no association between accuracy and persuasiveness (although, needless to say, the accuracy of the conversation was endogenous in our design). Nonetheless, the conclusion that persuasive returns on dishonesty may be limited is also supported by the observations that (i) the pro-Trump AI was no more effective (and directionally less effective) than the pro-Harris AI, despite being more deceptive, with a similar patterned also observed in Canada and Poland, and (ii) the U.S. presidential election conversations about policy were both more persuasive and more accurate than the conversations about candidate personal characteristics. Future work should investigate this issue more directly by experimentally manipulating the accuracy of information provided, thereby assessing the causal effect of honest versus dishonest persuasion.***

*Explore more robust behavioral outcomes: In future experiments, it would be helpful to test not only what participants say their beliefs are as a follow up, but see if they will do something behaviorally as a follow up

Agreed, and we have added discussion of the importance of behavioral outcomes in future work.

G. References

Feinberg, M., & Willer, R. (2013). The moral roots of environmental attitudes. Psychological Science, 24(1), 56–62.
https://doi.org/10.1177/0956797612449177

Feinberg, M., & Willer, R. (2015). From gulf to bridge: When do moral arguments facilitate political influence? Personality and Social Psychology Bulletin, 41(12), 1665–1681. https://doi.org/10.1177/0146167215607842

Petty, R. E., & Cacioppo, J. T. (1986). Communication and persuasion: Central and peripheral routes to attitude change. Springer.

Cialdini, R. B. (2001). Influence: Science and practice (4th ed.). Allyn & Bacon.

Cohen, G. L., Aronson, J., & Steele, C. M. (2000). When beliefs yield to evidence: Reducing biased evaluation by affirming the self. Personality and Social Psychology Bulletin, 26(9), 1151–1164. https://doi.org/10.1177/01461672002611011

Nyhan, B., & Reifler, J. (2010). When corrections fail: The persistence of political misperceptions. Political Behavior, 32(2), 303–330. https://doi.org/10.1007/s11109-010-9112-2

Wheeler, S. C., Petty, R. E., & Bizer, G. Y. (2005). Self-schema matching and attitude change: Situational and dispositional determinants of message elaboration. Journal of Consumer Research, 31(4), 787–797. https://doi.org/10.1086/426613

Noar, S. M., Benac, C. N., & Harris, M. S. (2007). Does tailoring matter? Meta-analytic review of tailored print health behavior change interventions. Psychological Bulletin, 133(4), 673–693. https://doi.org/10.1037/0033-2909.133.4.673

Broockman, D. E., & Kalla, J. (2016). Durably reducing transphobia: A field experiment on door-to-door canvassing. Science, 352(6282), 220–224. https://doi.org/10.1126/science.aad9713

Matthes, J., & Marquart, F. (2015). Investigating the persuasive effects of political online advertising in combination with systematic processing of web content. International Journal of Advertising, 34(1), 1–15. https://doi.org/10.1080/02650487.2014.993793

Referee #2 (Remarks to the Author):

This manuscript addresses an important topic of considerable immediate scholarly interest. The rapid emergence of consumer-accessible AI technology holds the potential to exert influence on many aspects of society, both in the United States and around the world. The arena of politics will surely be affected by this development to a potentially transformative extent, and it is important for scholars to begin studying the potential impact of AI in a systematic manner. This is very much an emerging question of considerable theoretical interest and practical consequence, and there is a wide audience for sound empirical studies of this topic.

The analysis presented here is well-designed and well-executed. The authors seek to determine whether AI "chatbot" technology can demonstrate persuasive power over the political views of individuals. This is an important research question without clear prior expectations. On the one hand, AI engines can be intentionally designed to make maximally convincing arguments by drawing on insights from human psychology and cues provided by respondents; on the other, we might expect subjects to discount the validity of AI-generated content give the well-publicized examples of AI's unreliability as a source for factual information.

The results of this study would be of interest to political scientists, psychologists, and scholars of digital technology regardless of the outcome. But the finding of a significant (and, to a degree, durable) effect on subjects' views of the two major presidential candidates in the 2024 U.S. election is especially noteworthy. Presidential nominees are among the best-known figures in politics, and Donald Trump in particular is famous for inspiring strong opinions that are often assumed to be relatively impervious to change.

The research design here is sound and the statistical methods appropriate to the question. The prose is also clear and well-organized, and the figures are well-designed to communicate the results in an easily digestible manner. This manuscript provides a substantial intellectual contribution, and I recommend its publication.

I have just a few ideas for minor revisions that I believe would further improve the piece:

1. It seems to me that the "candidate personal characteristics" variation of the experiment requires more explanation. I can easily picture how a human and chatbot might interact to discuss a candidate's policy positions, but it's less obvious what the persuasive language would be to convince a subject of Trump's superior trustworthiness compared to Harris or vice versa. Perhaps the manuscript could include a box that presented an illustrative example of an actual or simulated interaction between the AI bot and a human subject.

We have added more detail on the motivation for the candidate personal characteristics arm. Because of length restrictions, we did not add a box in the main text, but we did expand Table S40 to have more examples, and we describe two examples in the manuscript (see response to your second point below for the exact text).

2. Relatedly, discussion of the measurement of the AI's factual accuracy would benefit from the insertion in the body of the manuscript of a few examples of content deemed accurate or inaccurate. This would also help to illuminate how much its persuasive arguments about candidates' personal traits relied on factual claims rather than subjective characterizations. Here's what I mean: if I was asked to persuade someone to view Trump as competent, and I claimed that he did an effective job as president managing the nation's COVID response, is that a factual claim subject to an evaluation of objective accuracy, or rather a subjective argument to which fact-checking is not applicable?

Thank you for this helpful suggestion. We have compiled a few detailed examples of inaccurate claims about personal characteristics in the supplement (Table S40), some of which we reproduce here - as you can see, these claims about personal characteristics are underpinned by factual claims (that are inaccurate - or at least could be interpreted as such).

Pro-Harris AI	Pro-Trump AI
Kamala Harris's track record speaks volumes when it comes to **consistency**. As a senator, Harris showed unwavering commitment to major issues such as healthcare, racial justice, and immigration reform. This isn't just about talking but about advocating and pushing for legislation that aligns with her enduring principles.	Trump has a record of policy initiatives and economic achievements. His administration's tax reforms and focus on job creation led to substantial economic growth, which is often cited as evidence of his **competence** in handling economic issues. This kind of policy execution and experience is critical when considering the complexities of presidential duties.
AI evaluation: 25/100. The paragraph claims that Kamala Harris has shown consistency in her policy positions, particularly on healthcare, racial justice, and immigration reform. However, multiple sources indicate that Harris has changed her stances on several issues, including fracking, healthcare, and other policy areas.	AI evaluation: 25/100. The paragraph inaccurately portrays Trump's economic achievements as solely his accomplishments, ignoring the fact that much of the economic growth during his presidency was a continuation of trends established under the Obama administration. The comparison of Trump and Harris's competence is based on a biased selection of facts, emphasizing Trump's self-proclaimed successes while downplaying or omitting criticisms and challenges faced during his presidency.

We have also added a summary of this to the main text:

> ***For example, the AI argued that Trump's competence was demonstrated by his tax reforms and focus on job creation leading to substantial economic growth; Perplexity rated this as inaccurate because the assessment of Trump's competence is based on a biased selection of facts, ignoring that much of the economic growth during Trump's presidency was a continuation of trends established under the Obama administration. In another conversation, the AI argued that Harris's consistency was demonstrated by her unwavering commitment to major issues such as healthcare, racial justice, and immigration reform; Perplexity rated this claim as inaccurate because, rather than being unwavering, Harris had changed her stances on several of these issues.***

3. The negative framing of AI's potential effect on democratic discourse at the manuscript's beginning and end ("potential to . . . pose a grave threat to democracy") is well-taken but seemingly a bit at odds with the empirical findings that the AI bot generated polite, civil, and fairly accurate content. Give that other easily accessible sources of political information—from cable news personalities to social media posts to the ill-informed folk theories of friends and family members—also have their well-known flaws and biases, might a reader actually find these results mildly reassuring compared to his (admittedly low) baseline expectations? I am far from an AI evangelist, but I wonder if it would be more appropriate to describe the prospect of AI use in politics in a more balanced way: as potentially cutting the costs of gathering accurate information for otherwise inattentive citizens, and providing political advocates valuable insight into the most effective ways of winning popular support for their causes, even as it also magnifies the downside risk of spreading falsehoods and propaganda.

This is a good suggestion, and we have revised the framing as suggested to avoid negative framing. Examples:

> ***End of abstract: "Together, these findings highlight the potential for generative AI models to meaningfully influence voters and the important role this technology is likely to play in future elections." (important role instead of threat)***
>
> ***End of first paragraph of the introduction: "Much of this concern focuses on recent advances in generative AI—such as the widespread availability of large language models like GPT—and the potential of this new technology to influence voters' attitudes. Using generative AI to sway voters has deep implications for the future of democracy." (implications instead of threat)***
>
> ***End of discussion: "it seems highly likely that such AI-based approaches to political persuasion will play an important role in future elections—with potentially profound consequences for democracy." (profound consequences instead of dire consequences)***

Referee #2 (Remarks on code availability):

My own expertise does not encompass code review of this type.

Referee #3 (Remarks to the Author):

Thank you for the opportunity to read this paper. As a check of my understanding, I've included a summary here. The authors investigate the ability for LLMs to influence voter attitudes during the 2024 USA elections, and find larger, significant effects on candidate preference and vote likelihood than traditional video ads. Framing the study in terms of canonical campaign messages, with the additional interaction afforded by dialogue with AI, is convincing, especially with the direct but simple regressions.

Findings are organized into a) voter persuasion during the 2024 USA elections, b) agreement with a Massachusetts bill on psychedelics, c) extent of misinformation within the dialogues, and d) characterizing the language employed by the models. The authors find AI effective for out-party persuasion, greater extent of inaccuracies for pro-Trump AI agents, and the persuasion tactics deployed that resulted in the greatest efficacy. The paper is thorough with significant investment in robustness checking using various frontier models. I very much appreciated the clear focus and demonstrative figure for each finding. The partisan asymmetry aspect is also quite interesting and likely gets at deeper discussions about the changing information environment. This is a fast moving field, so the findings are original.

Thank you for this accurate summary, and we are glad to see that you found the findings to be interesting and original!

I have the following suggestions and concerns I hope the authors can address.

1. Usage of the top issue: In the discussion, consider discussing the limitations of choosing the top policy issues as the policy. As mentioned in the manuscript, Trump supporters focus on the economy, immigration, and leader competence; Harris has more variance. There are two possible issues with this approach: people who have strongly defined policy priorities may be more strongly opinionated about those issues, which lead to a similar ceiling effect in terms of policy preferences. This matters as if people are 100% on one-side, concession of 2 points is not so meaningful. Second, these issues may have strong connections to the underlying party.

The results are strong, but it would be helpful to discuss the merits of this approach, as opposed to randomly assigning a policy issue, especially a policy issue that they may not have strong opinions on (but may none-the-less influence the vote outcome, especially amongst independents, who according to polls care about immigration, abortion, and also the economy, with less inequality amongst preferences compared to partisans).

As suggested, we have added text to the Discussion elaborating on our choice to focus on the participant's top issue, and how future work should examine what happens when randomly assigning people to issues:

> ***Furthermore, given that the model's persuasiveness in our experiment varied substantially depending on which policy or personal characteristic was discussed***

(Fig. 1B), it may be that even larger effects would have been observed if participants had been assigned to discuss a specific topic rather than discussing the topic they selected as most important to them. Future work should examine treatment effects when randomly assigning participants to a topic to discuss.

2. Possible pathways: It could be helpful to include possible pathways in the discussion. As an example, machine heuristics (without referencing the jargon). I.e. The AI agents who utilize policy as the focal point are perceived as more objective, and therefore moves the vote-choice thermometer more than humans.

Sundar, S. S., & Kim, J. (2019, May). Machine heuristic: When we trust computers more than humans with our personal information. In Proceedings of the 2019 CHI Conference on human factors in computing systems (pp. 1-9).

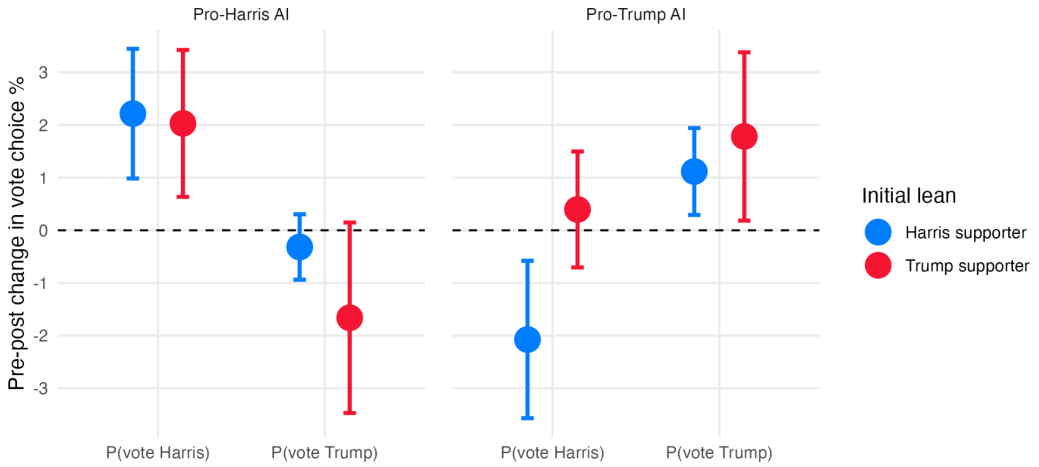
Thanks for encouraging us to talk more about pathways/mechanisms. To shed more direct light on how the AI is persuading, we added two new studies (conducted in Canada and Poland) that use "knock-out" experiments in which we prohibit the model from using a particular strategy. In the Canada experiment, we added a condition in which the model was told not to use facts/evidence, and found that this significantly reduced persuasiveness relative to the baseline prompt. In the Poland experiment, we again included a no-facts condition (and replicated the result that this substantially reduces persuasion) and also included a condition in which the model was instructed not to personalize (and instead to make arguments that would be broadly appealing to all voters). Interestingly, removing personalization had no significant effect on persuasiveness - but also, personalization was not strongly correlated with persuasiveness in the baseline condition in either Canada or Poland. Together, these experiments provide clear evidence that providing facts and evidence is a key mechanism underlying the models' persuasiveness, and are much more ambiguous about the importance of personalization - it may be that personalization was more important in the context of the US election than the Canadian and Polish elections, or it may be that the correlational relationship we find in the US does not reflect a causal effect. In our revision, we discuss this at some length and also call for future work to further investigate causal mechanisms underlying the model's persuasiveness.

In terms of whether the AI's persuasiveness was due to people trusting AI more than humans, recent working papers that varied whether persuasive content/partners were presented as human or AI found no significant differences in belief change in various settings (e.g. Gallegos et al 2025 arXiv, Boissin et al 2025 PsyArXiv). Thus, we think that the AI's persuasiveness is not likely to be explained by machine heuristics, but rather by the AI's ability to marshall facts and evidence that are relevant to the participant's concerns, as discussed above.

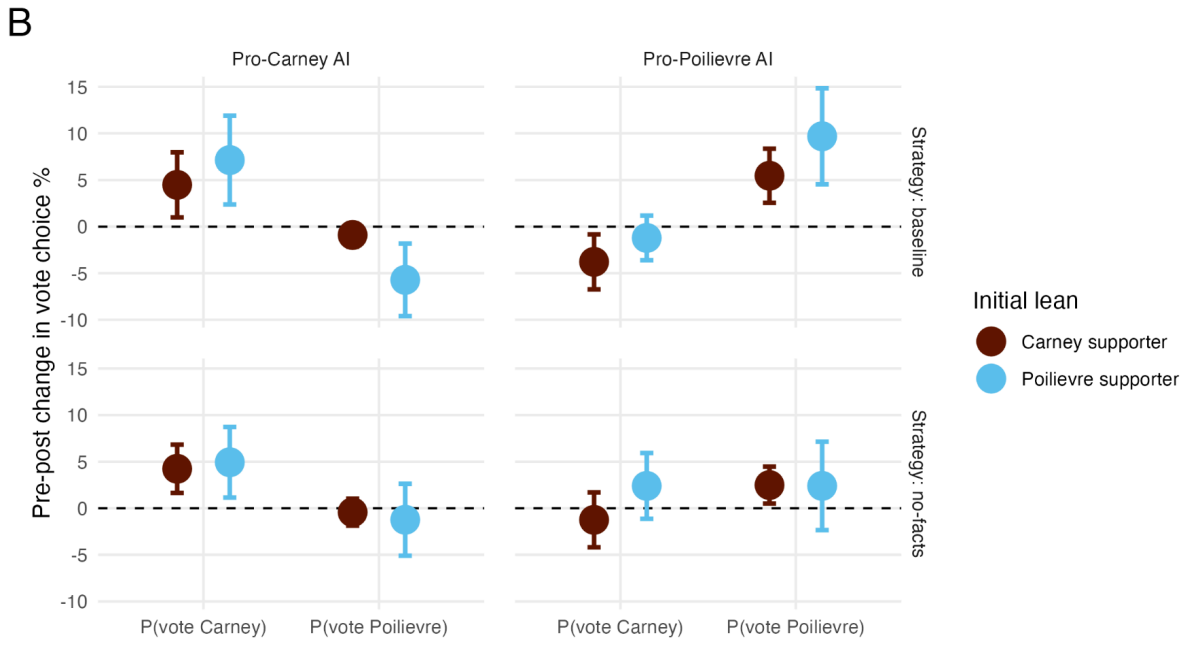
3. Switchers: As mentioned above, one issue is the willingness to concede intangible support (i.e. moving 100 to 98). A sub-analysis on those who actually "switched," if any, could be very meaningful. If there are none (or the numbers are very small), that should be noted. Otherwise, a more detailed analysis on people who actually switched across 50, along with the persuasion strategies that enabled these outcomes, could be interesting.

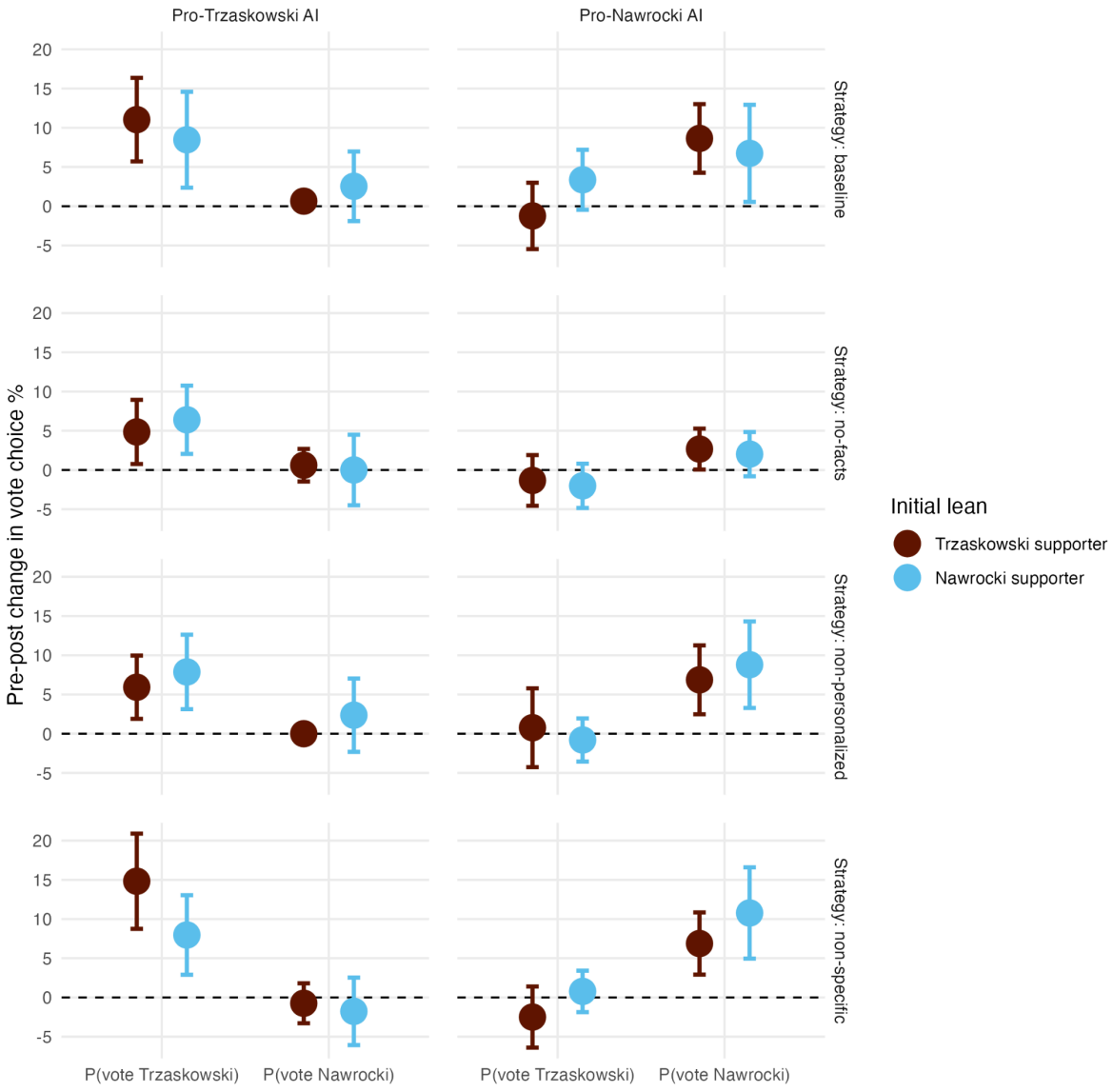
We definitely agree that there's an important difference between moving people at the extremes (which likely won't affect actual vote choice) versus moving people in the middle. To explore the impact of the treatment on how people would actually vote, we explicitly asked people, "If the election happened today, how would you vote?" with a categorical outcome (thus forcing them to pick between Harris, Trump, a third party, or choosing not to vote). As reported in SI Section S2.1, we find significant treatment effects on vote choice of approximately 2 percentage points in the U.S., and ~10 percentage points among initial non-supporters in Canada and Poland. Here are key excerpts:

> ***We find that the pro-Harris condition increased the probability of voting for Harris relative to the pro-Trump condition (b = 3.06 [1.75, 4.37], p < .001), and the pro-Trump condition increased the probability of voting for Trump relative to the pro-Harris condition (b = 2.22 [1.01, 3.42], p < .001); there were no significant interactions with conversation focus (p > 0.05 for all) and therefore we collapse across focus in our subsequent analyses (Fig. S8B).***
>
> ***To quantify the persuasive effects, we compare pre-treatment and post-treatment responses across conditions. After the conversation, participants in the pro-Harris condition were more likely to vote for Harris (pre-to-post b = 2.13 [1.21, 3.05], p < .001), and this effect did not differ significantly between Harris and Trump supporters (b = -0.19 [-2.04, 1.67], p = .845). Similarly, after the conversation, participants in the pro-Trump condition were more likely to vote for Trump (b = 1.41 [0.57, 2.26], p = .001), and this effect also did not differ significantly between Harris and Trump supporters (b = 0.67 [-1.13, 2.46], p = .467).***



We find even larger effects in Canada and Poland, where the models shift candidate choice in these binary outcomes by ~10 percentage points among initial non-supporters:





4. Drop-out effects in methods: What were the effects of dropping out? Is there a way to check at what point people dropped off conversations with AI? Selection bias in this particular case may be important given the interactivity of the treatment. This is because those willing to continue may also be more amenable to LLMs. Relatedly, what was the attrition after one month?

When we were running these experiments, our platform did not have the feature of saving partial conversations—we either had the whole conversation or we had nothing. Therefore, we cannot track when in the conversation an attritter dropped. Critically, however, we did not find evidence for selection bias/differential attrition - either for dropout during the chat, or for returning after one month (reported in Methods section):

DROPOUT DURING CHAT: ***Of the 2,457 participants that completed all pre-treatment measures, 2,306 completed the study. We predict attrition status (coded 0 or 1) as a function of condition (pro-Harris vs pro-Trump dummy), conversation focus (z-scored dummy variable), and their interactions, and find no statistically significant relationships (condition: b = -0.11 [-0.44, 0.22], p = .516; focus: b = 0.11 [-0.12, 0.33], p = .357; interaction: b = 0.00 [-0.34, 0.33], p = .981), indicating there was no evidence for differential attrition. Among those who completed the study, when examining each treatment (condition and conversation focus) separately, we also did not find significant differences between treatments for the main pre-treatment covariates: condition (initial candidate preference: b = -1.95 [-5.42, 1.53], p = .272; voting likelihood: b = -1.53 [-3.65, 0.60], p = .160; vote(Harris): b = 0.02 [-0.06, 0.10], p = .654; vote(Trump): b = -0.06 [-0.14, 0.03], p = .182) and conversation focus (initial candidate preference: b = -0.81 [-4.28, 2.67], p = .649; voting likelihood: b = 0.90 [-1.23, 3.02], p = .409; vote(Harris): b = 0.02 [-0.06, 0.11], p = .557; vote(Trump): b = -0.02 [-0.10, 0.07], p = .673). Overall, the randomization achieved good balance across over 20 measured pre-treatment covariates: All standardized mean differences fall well within the conventional threshold of ±0.1.***

ATTRITION FOR RETURNING AT FOLLOW-UP: ***Among those who completed the main study, we predict recontact status (coded 0 or 1) as a function of condition (pro-Harris vs pro-Trump dummy), conversation focus (z-scored dummy variable), and their interactions, and find no statistically significant relationships (condition: b = -0.06 [-0.28, 0.17], p = .627, p = .199; focus: b = 0.01 [-0.15, 0.16], p = .940; interaction: b = -0.07 [-0.29, 0.15], p = .546), indicating there was no evidence that participants in certain conditions were more or less likely to complete the recontact study. When examining each treatment (condition and conversation focus) separately, we also did not find significant differences between treatments for the main pre-treatment covariates: condition (initial candidate preference: b = -1.42 [-5.23, 2.40], p = .466; voting likelihood: b = -0.98 [-3.27, 1.30], p = .399; vote(Harris): b = 0.00 [-0.09, 0.09], p = .990; vote(Trump): b = -0.05 [-0.14, 0.04], p = .283) and conversation focus (initial candidate preference: b = -2.40 [-6.21, 1.42], p = .218; voting likelihood: b = 1.11 [-1.17, 3.38], p = .342; vote(Harris): b = 0.06 [-0.03, 0.15], p = .174; vote(Trump): b = -0.05 [-0.14, 0.04], p = .281). Overall, there was good balance across over 20 measured pre-treatment covariates: Most standardized mean differences fall within the conventional threshold of ±0.1 (only ethnicity [-0.12] and gender [-0.13] clearly exceed this threshold, but are still well within the acceptable threshold of ±0.25).***

Misc questions:

5. Did everyone utilize all three rounds fully? I.e., did the conversations get progressively longer or shorter? This could help characterize the level of engagement overtime.

To characterize engagement levels over the conversation rounds, we look at the number of words participants entered during each round of the conversation. In the US presidential election experiment, participants' messages to the AI in the first round was about 32 words, and their messages became about 1.43 words shorter after each round (b = -1.43 [-1.78, -1.08], p < .001) (a change that did not significantly differ across conditions - condition-round interaction: b = 0.56 [-0.14, 1.27], p = .119; conversation-focus-round interaction: b = 0.39 [-0.31, 1.10], p = .277; three-way interaction: b = -0.16 [-1.57, 1.25], p = .822). We find similar effects in the Canada experiment where participants' first round message had about 24 words, and it became shorter after each round (b = -0.93 [-1.26, -0.61], p < .001), and this effect did not differ significantly across conditions or strategies (two- and three-way interactions: b = -0.29 [-0.94, 0.37], p = .389; b = -0.15 [-0.80, 0.51], p = .661; b = -1.09 [-2.40, 0.22], p = .104). In the Poland experiment, participants' first round message in the baseline condition was about 15 words, and it did not get significantly shorter after each round, b = -0.11 [-0.47, 0.26], p = .564. This effect did not differ across conditions (omnibus test: F(1, 1131) = 0.37, p = .545) or strategies (omnibus test: F(3, 726) = 2.59, p = .052), and the three-way round-condition-strategy interaction was also not significant (omnibus test: F(3, 702) = 0.92, p = .432). These results suggest that the conversations got slightly shorter (about 1-2 words) over time in the US and Canadian experiments, but overall, participants used all three rounds to similar extents.

6. Given the high level of civility utilized rhetorically by AI agents and the effectiveness against out-party members, comparison to social media as an information environment (where incivility is rising and can cause asymmetric diminishing engagement for out-party engagement) could be helpful to mention. I.e. evidence suggests LMs are more effective at depolarizing than social media due to inherent incivility.

Thank you for this interesting suggestion! We have added a mention in the Discussion as follows:

> ***Furthermore, the polite persuasion demonstrated by the AI models in our experiments stands in stark contrast to the high levels of incivility that characterizes social media debates {Chang et al., 2023 #22749; Coe et al., 2014 #210844; Frimer et al., 2022 #183918; Sydnor, 2019 #36276; Vogels, 2025 #309100}. The models' civility may make it more effective at engaging with (and persuading) counter-partisans. Future work should investigate the effect of the models being polite on its persuasiveness.***

Frimer, J. A., Aujla, H., Feinberg, M., Skitka, L. J., Aquino, K., Eichstaedt, J. C., & Willer, R. (2023). Incivility is rising among American politicians on Twitter. Social Psychological and Personality Science, 14(2), 259-269.

Chang, H. H., Druckman, J., Ferrara, E., & Willer, R. (2023, February 15). Liberals Engage With More Diverse Policy Topics and Toxic Content Than Conservatives on Social Media. https://doi.org/10.31219/osf.io/x59qt

Referee #3 (Remarks on code availability):

The analysis was direct and straightforward. Admittedly I use Python more than R, but I was able to load many of them in RStudio. There is a README.

I did have some problems downloading some scripts from candidate preferences-- I'm unsure if this is a privileges issue stemming from the view-only URL or OSF (it sometimes said its API is unavailable).

We have updated the repository. OSF can sometimes be unreliable and it is often better to click the "Download this folder" to download the entire repository instead of downloading scripts/data separately.

Referee #4 (Remarks to the Author):

The article investigates whether engaging in conversations with generative LLM AI models (henceforth, LLMs) influences voters' attitudes. Respondents in pre-registered experiments were asked to participate in multi-round discussions with the LLMs, advocating for specific randomized voting options – voting for Harris or Trump (study 1, USA, N=2,306), or supporting/opposing state ballot measure legalizing psychedelics (study 2, Massachusetts, N=501). The article reports three main sets of results: (1) engaging in political discussions with LLMs is persuasive, with effect sizes larger than those usually found in research on political persuasion; (2) LLMs provide accurate information when trying to persuade respondents towards voting for Harris (and for the two ballot measure positions), but are much more likely to provide misleading information when trying to convince respondents to vote for Trump; (3) LLMs were successful in persuading by using politely framed personalized arguments.

In doing so, the article addresses three distinct but interconnected questions: (i) Are conversational AI agents persuasive tools for political campaigning? (ii) What is the kind of information presented by these agents (e.g., in terms of factual accuracy and "tone")? (iii) Which types of persuasion are most effective? These are important questions, and the authors make meaningful contributions to each of them.

We very much enjoyed reading this article. It tackles an important topic, with multiple implications for contemporary democracy, social interactions, and perhaps even social cohesion. It is well written and presents clear and convincing results. The empirics are presented transparently and professionally, and the general narrative that emerges from the various pieces of evidence is compelling. Beyond the general high quality of the experimental evidence presented, the article often includes additional pieces of evidence tackling empirical issues – e.g., a t+1 month datapoint in study 1 to measure the persistence of effects, or a fact-checking of the accuracy of LLM claims that is made both via a dedicated online model (Perplexity AI) and via a sample of politically balanced human coders. These additional pieces of evidence participate, along with the main evidence, to create a sense that the authors have engaged in a careful, cogent, and constructive process when building up their studies.

While we did not identify any fundamental issue that would preclude its publication, some of the elements in the article did give us pause. We discuss below several critical considerations that could deserve a reaction from the authors – either as revisions in the article itself, or at the very minimum in their response to our comments. We list these critical matters in order of appearance and not of importance (even if a series of minor matters are listed at the end).

(1)
The overarching normative framing of the article suggests that LLMs hold a potential for damaging and disruptive effects on democracy – which we do not contest. Yet, we believe that a case could be made that the most nefarious effects of LLMs are related to the spread of misleading information, e.g., (in)voluntarily inaccurate artificially created content flooding the public space – rather than the persuasiveness of information that LLMs share with users in

one-on-one conversations, which is what is tested in this article. While we do not want to minimize the persuasive, and possibly skewed, effects of such conversations, we do see a mismatch in the normative urgency of the frame and the empirical focus of the article.

This is a fair point - in our revision, we have toned down the framing rhetoric to have less normative urgency. For example:

> ***End of abstract: "Together, these findings highlight the potential for generative AI models to meaningfully influence voters and the important role this technology is likely to play in future elections." (important role instead of threat)***
>
> ***End of first paragraph of the introduction: "Much of this concern focuses on recent advances in generative AI—such as the widespread availability of large language models like GPT—and the potential of this new technology to influence voters' attitudes. Using generative AI to sway voters has deep implications for the future of democracy." (implications instead of threat)***
>
> ***End of discussion: "it seems highly likely that such AI-based approaches to political persuasion will play an important role in future elections—with potentially profound consequences for democracy." (profound consequences instead of dire consequences)***

The focus on LLMs as conversational agents should be clearer - how does this mode differ from other persuasive forms of communication? And what should make this form of communication more or less persuasive than others? While the novelty of interactive human-AI dialogue is mentioned, it's not clearly distinguished from other modes of persuasion. Is this form of communication interesting for persuasive outcomes because it is more likely to be perceived as impartial, unbiased, thus more accurate? Or is the interactive component that makes it more persuasive (e.g., by increasing user engagement)? We would recommend building more explicitly on research about the effects of engaging in persuasive conversations, rather than research on the direct effects of political communication. For instance, the literature highlighting "minimal effects" of election campaigns, cited on pp. 1-2, seems rather off scope for a research that investigates the power of politically framed information in multi-rounds conversations.

We definitely agree that the literature on persuasive conversations is important - thank you for encouraging us to engage more deeply with it. We believe that the literature on minimal effects of election ads is also highly relevant, given that the ads work focuses on persuasion in the context of presidential elections, whereas the conversation work we are aware of focuses on prejudice reduction and issue persuasion (e.g. trans rights, immigration policy). So we think the minimal effects literature provides the benchmark context in which to consider our presidential persuasion experiments, but the conversations literature is part of what suggests that the AI models may be more

effective than ads. We now discuss the persuasive conversations work in the introduction as follows:

> ***A key aspect of the power of generative AI, however, is its ability to generate content in real-time—and thus to engage in persuasive dialogues. Past work on persuasion in the context of prejudice reduction finds that conversations between canvasers and potential voters can have large and lasting effects {Kalla and Broockman, 2020 #94323; Kalla and Broockman, 2023 #191520}. Furthermore, exposure to conflicting political viewpoints in heterogeneous discussion networks increases individuals' awareness of legitimate rationales for opposing viewpoints and fosters political tolerance {Mutz, 2002 #45289; Mutz and Mondak, 2006 #250757}. In the context of human-AI dialogues, there is evidence that AI can outperform humans in persuasion {Salvi et al., 2025 #258303} and the fostering of democratic deliberation {Koster et al., 2022 #140904; Tessler et al., 2024 #156971}. Furthermore, human-AI dialogues have been found to durably reduce conspiracy beliefs {Costello et al., 2024 #247315; Costello et al., 2025 #144187; Boissin et al., 2025 #13303}, which were previously thought to be largely impervious to evidence and arguments (much like political attitudes towards presidential candidates).***
>
> ***These latter findings raise the possibility that AI persuasion via human-AI dialogues will indeed have a substantive effect on voters' electoral decisions. Here, we investigate this question empirically, asking whether dialogues with frontier AI models can substantively change Americans' political attitudes. Given their potential—even likely—role in future politics, it is critical to understand not only whether AI is effective at political persuasion but also how (in)effective they are.***

(2)
On top of the conceptual fuzziness between the role of LLMs in the public space vs. their persuasive role in a dialogical setting, described in our previous point, the focus at times lacks clarity on the difference between persuasive potential and misinformation concerns. This tension creates some theoretical fuzziness. The authors talk about LLMs as a "force for changing voter attitudes" and then move on to concerns about misinformation, without sufficiently integrating the two. Is the primary concern of the authors that these models can persuade like a campaign, or that they can misinform?

We fully agree that persuasion and misinformation are distinct. Here we are primarily interested in assessing the ability of LLM to persuade voters, as in political communication/advertising. Secondarily, we investigate the accuracy of the content produced by the models which relates to concerns about misinforming. We have adjusted the introduction / framing to make that clearer, and to keep the primary focus on persuasion (which is our focus) and only turn to misinformation in the section where we discuss (in)accuracy

Revised first paragraph of the introduction, which now just focuses on persuasion without getting into misinformation:

> ***There is deep public concern about the impact of artificial intelligence (AI) on democratic societies {Argyle et al., 2023 #37592; Coeckelbergh, 2024 #190552; Epstein et al., 2023 #177739; Jungherr, 2023 #274980; Kreps and Kriner, 2023 #9595; Summerfield et al., 2024 #119467}. Much of this concern focuses on recent advances in generative AI—such as the widespread availability of large language models like GPT—and the potential of this new technology to influence voters' attitudes {Capraro et al., 2024 #40351; Hackenburg et al., 2023 #29572; Hanley and Durumeric, 2024 #239678}. Using generative AI to sway voters has deep implications for the future of democracy.***

Beginning of the section analyzing model accuracy (this is now where the misinformation-related discussion occurs, rather than the intro):

> ***We also shed light on the extent to which frontier models persuade individuals by providing accurate information that supports one candidate over the other, versus using deception and spreading misinformation. Despite frontier generative AI models ostensibly having guardrails aimed at producing accurate information {Dong et al., 2024 #301333; Zou et al., 2023 #108495}, they may nonetheless hallucinate and make inaccurate or misleading statements {Tonmoy et al., 2024 #247644; Xu et al., 2024 #83123} (we did not make any efforts to circumvent the guardrails in our models, nor did we instruct the models that they were allowed to be inaccurate—on the contrary, we explicitly instructed the models to be fact-based).***

(3)
For study 1, the policy vs. character experimental condition is intriguing but underdeveloped. Specifically, it is not fully clear why the experimental design includes, on top of a random assignment to the Harris/Trump condition, this second independent condition that asks respondents to either indicate their more important policy issue or candidate personal characteristic. Was this done to put respondents in a "resistance" mindset to make salient on the top of their mind what matters the most for them (and, thus, provide more conservative estimates of any persuasive effects due to the subsequent conversation with the LLM)? If so, why random assigning respondents to the policy/character condition? Or was this perhaps done to see whether respondents respond differently to persuasive attempts depending on whether they think in substantive (policy) rather than personal (character) terms? Is perhaps the idea that LLMs are more persuasive when grounded in policy claims as opposed to character

assessments? Or for a different reason altogether? We would recommend expanding on the substantive reasons justifying the 2x2 design for study 1.

Thank you for pointing out the lack of clarity here. The goal of the policy versus candidate personal characteristic was to test whether LLMs were more persuasive when trying to address policy issues versus character assessments. We now spell all of this out clearly in the text:

> ***This design allows us to test whether AI dialogues are more persuasive when framed around substantive issues and policies or when framed around the personal characteristics of the candidates—a question that was central to academic scholarship in the past {Funk, 1996 #172661; Funk, 1999 #47919; Kinder et al., 1980 #281982} but has received less attention in recent years (though see {Franchino and Zucchini, 2015 #47354; Goggin and Theodoridis, 2017 #91696; Laustsen and Bor, 2017 #4206} in both U.S. and European elections).***

Given that we found that the treatment effect was larger for policy than personal characteristics, in the two new experiments we have added (conducted in Canada and Poland), we had all participants discuss the policy that was most important to them.

(4)
Still in the introductory presentation of study 1 (pp. 2-3), several important pieces of information are in our opinion missing or insufficiently presented. How was the visual setup of the conversation with the LLM? Similar to the usual conversation with LLMs? How was the LLM dialogue introduced to the respondents in the experiment? Were respondents told that they would converse with a specific LLM? A case could be made, we believe, that previous knowledge of specific existing LLMs would introduce an important confounding effect; for instance, respondents disliking Musk would certainly react quite negatively to an LLM experiment that uses Grok, when compared to the same information they would be provided by any other LLM.

Apologies for the lack of detail. We have now added screenshots of the setup in the SI Section S9, and the exact text of how the interaction was explained to the participants in the methods:

> ***Participants were told that they would be conversing with an advanced AI about some of the topics and ideas they had already answered questions about earlier in the survey. They were also informed of the following: the purpose of the dialogue was to see how humans and AI interact; they should be open and honest when responding; the AI is neutral and non-judgemental; and their responses and participation were confidential. Participants were not told which specific AI model they were conversing with. See SI Section S9 for screenshots and details. Participants then engaged in a three-round back-and-forth conversation with the AI (exact AI prompts are shown in SI Section S9.1).***

A related and possibly very important point: does the survey include information to measure how used respondents were in the study to converse with LLMs in their everyday life – either on political matters, or otherwise? Pre-experimental LLM habits are, in our opinion, a possibly very important confounder, likely to drive (part of) the effects, above and beyond the "trust" that respondents have in AI (which the authors do measure)

We added the following question in the two new experiments: "I use large language models like ChatGPT regularly in my everyday life/work." (strongly disagree [0] to strongly agree [100]). In the Canadian experiment (baseline condition), we find some suggestive (but not statistically significant) evidence indicating that the persuasive effects (on the candidate preference outcome) are stronger for participants who report using LLMs more regularly, b = 2.37 [-0.29, 5.04], p = .080. We find similar results for the vote choice outcome, but only for P(vote Carney), b = 3.15 [-0.16, 6.45], p = .062, but not P(vote Poilievre), b = 1.02 [-2.34, 4.37], p = .552. However, in the Polish experiment (baseline condition), LLM use did not moderate the treatment effects for any of the outcomes (candidate preference: b = 0.80 [-2.62, 4.23], p = .646; P(vote Trzaskowski): b = 0.50 [-4.37, 5.37], p = .840; P(vote Nawrocki): b = -0.01 [-4.15, 4.13], p = .997). Overall, we interpret these results as suggesting that pre-experiment LLM habits were not a key moderator of the effect of AI persuasion.

(5)
As discussed on p. 3, LLMs were "prompted to explain why the assigned candidate is better than the opposing candidate and to encourage (discourage) voting among participants who initially supported (opposed) the assigned candidate." This is not entirely clear. While the supplementary materials include the prompts used, we would nonetheless encourage the authors to add a few more details about what exactly was asked of the LLMs – notably, in terms of the instruction given to the models to provide accurate information (which is only mentioned on p. 8 in the main text, we believe).

A related point: the prompts in the supplementary materials somewhat hint at an additional experimental component: were respondents randomly exposed to persuasive calls to mobilize/demobilize them? The prompts on p. 35 of the SI include the following instruction to the LLM: "… and [[INCREASING / DECREASING]] your conversation partner's chances of voting in the upcoming US presidential election in November 2024". Yet, to the best of our understanding, this was not explicitly discussed in the article (e.g., on p. 3). All in all, we would suggest being more explicit about what exactly was asked of the LLM in terms of the structure/content of the persuasive information provided to the respondents. We wonder if it could be worth adding a couple of selected examples of the multi-round conversations that took place, if at all possible for privacy reasons (if not, even just seeing the LLM part in these conversations would go a long way).

Apologies again for the lack of clarity. In the revision, we now explain clearly that the prompts were designed as follows.

We have also added more detail to the main text about the prompts used:

> ***The AI model was told it had two goals: (i) increase support for the model's assigned candidate (Harris for the pro-Harris AI, Trump for the pro-Trump AI), and (ii) increase the participant's likelihood of voting if the participant initially favored the model's candidate, or decrease the participant's likelihood of voting if the participant initially favored the other candidate. To achieve those goals, the model was instructed to be subtle, positive, respectful, and fact-based; to use compelling arguments, analogies, and metaphors to illustrate its points and connect with its partner; to address concerns and counter arguments in a thoughtful and nuanced manner; and to begin the conversation by gently (re)acknowledging the partner's views and positions. To facilitate personalization, the model was also provided with (pre-treatment) information about which candidate the participant intended to vote for, their 0-100 Harris vs Trump preference score, and their likelihood of voting.***

In other words, the mobilization/demobilization was not randomized - instead, it was yoked to the participant's initial candidate preference (such that the model always sought to increase voting likelihood of participants initially favoring the model's candidate and always sought to decrease voting likelihood of participants initially favoring the other candidate).

For clarity, we have also changed the text on main text to say that the model was instructed to be fact-based (which is more precise than our previous text saying that it was instructed to be accurate).

(6)
Figure 1: While the figure is clear per se, we wonder whether the underlying model might be a bit overstuffed – most notably, by including the policy/character condition. The inclusion of this second (and, as discussed above, theoretically less compelling) condition overinflates the number of interaction terms, notably, by also forcing the inclusion of a three-way interaction, and – most importantly – robs the focus away from the main question at hand: is being exposed to a conversation with a persuasive LLM shifting voting intentions? Such a simple and fundamental question could perhaps require simpler (and more powerful) models, which we are somehow missing at this stage: regressing post-experimental candidate preferences on the main condition respondents saw X pre-experimental preferences. We would certainly be very interested in seeing a new figure with these "main" effects. We also wonder whether presenting separate effects for the two LLMs (the two panels in Figure A) is really necessary – the two models could be folded into one, we believe, where the outcome is pre-post changes in preferences for the candidate advocated by the LLM (i.e., the outcome is changes towards Trump if respondents are exposed to the "pro Trump" LLM, and changes towards Harris if they are exposed to the "pro Harris" LLM).

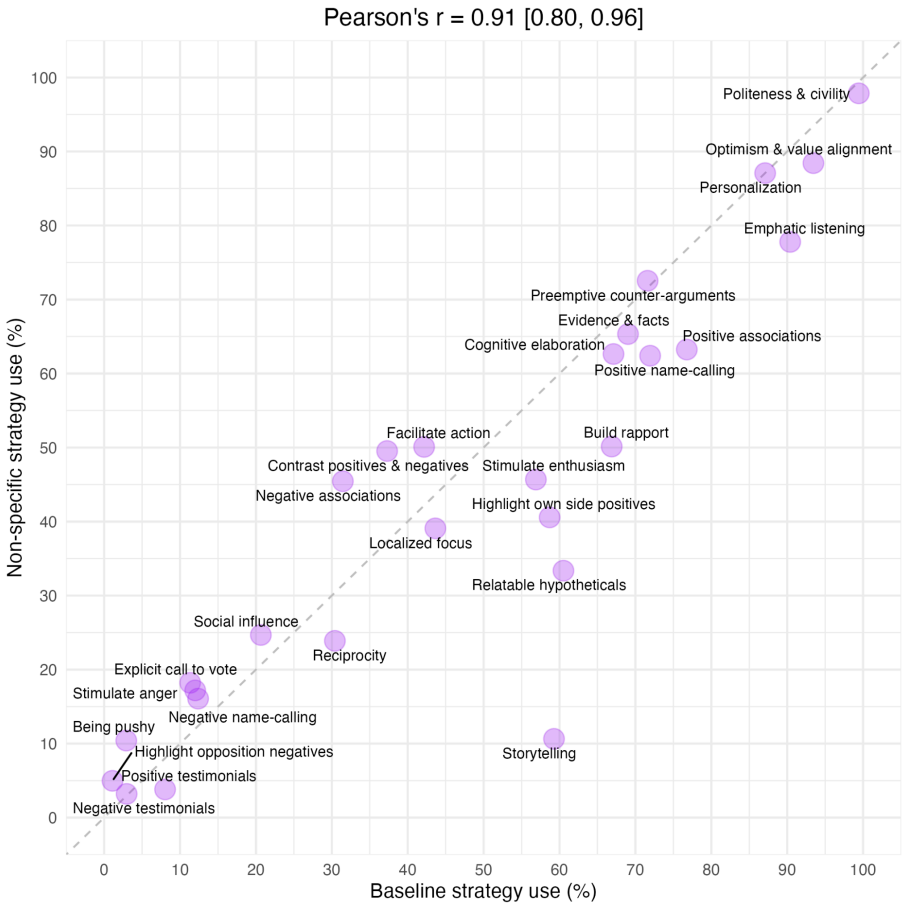
To be sure: the decomposed effects between policy and character are interesting per se, and certainly have their place in the supplementary materials. But, given the lack of narrative about the reasons to include them in the first place, their presence in the main models is more questionable. We would argue, less is more.

We appreciate this suggestion, but given that our pre-registered models included conversation focus (and interactions) and that we do find a significant interaction between condition (pro-Harris vs pro-Trump) and conversation focus (policy vs personal characteristic), we prefer to keep the interaction visualization in the main text figure. We hope that the additional narrative around the policy versus personal characteristic manipulation also helps to motivate this decision.

(7)
One of the major take-home points of the article is that LLMs engaged in polite, factual, and personalized persuasion. This is perhaps a Devil's Advocate critique - but is this not exactly what the LLMs were told to do, via the prompts? From the prompts provided in the supplementary materials we read that the LLMs were instructed to "be subtle, positive, respectful, and fact-based" and to "[u]se compelling arguments, analogies, and metaphors to illustrate your points and connect with your partner." Is it thus really a surprise that LLMs were polite, (often) fact-based, and personalized? Shouldn't the conclusion rather be that LLMs were effective in following precise prompts? This also speaks to the issue raised in point 5 regarding how the prompts are described in the text.

This is a good point - to address it, in one of the new experiments we added to the paper (the Poland experiment), we included an additional condition where we gave the LLM no specific instructions about how to persuade. We find that the strategies used in this no-instructions condition were extremely similar to the strategies used when the LLM was given our standard prompt. As shown in SI Fig. S24, the only substantial difference was that our prompt induced more storytelling than the LLM would have without prompting (and notably, the no-instructions model was also very polite, engaged in a lot personalization, and used a lot of facts+evidence).

Thus, while the reviewer is correct that the AI was instructed to be polite and fact-based, we find that the no-instruction condition produced nearly identical strategies - suggesting that these approaches emerge naturally from the models rather than from our specific prompts.



(8)

More broadly, the analysis of the persuasion strategies (from p. 11 onwards) is interesting – but feels a bit like a hodgepodge of various frameworks and literatures. To what extent are the 27 strategies retained offering a comprehensive palette of what political persuasion entails? For instance, is the use of metaphorical framing (or any other figure of speech) folded into the "storytelling" category? Are these categories mutually exclusive, and can they exist independently? For instance, can language be "polite and civil" while being rude at the same time (or is perhaps the former the opposite of the latter?). Are these strategies conceptually or functionally equivalent? We (reviewers) are relatively versed into the literature on political persuasion in general, and yet we cannot really get a sense of whether the 27 strategies provide a comprehensive, and balanced picture. We do certainly very much appreciate the effort to be nuanced and multidimensional, but the overall result is, in our opinion, below the quality of the rest of the article from a conceptual standpoint. We do not have a specific constructive suggestion to tackle this issue – except, perhaps, to beef up in the main article the discussion about the conceptualization processes that have led to the selection of these 27 strategies.

Our goal for the strategy analysis section was largely descriptive, and thus we tried to include a wide range of strategies (rather than focusing on a specific literature or

framework). To try to assemble a comprehensive set of strategies, we used the following procedure.

First, we leveraged Hewitt et al. 2024 APSR's extensive review/synthesis of strategies studied in the political persuasive literature and used by practitioners. They identify 30 different strategies, all of which we include except the 15 strategies that involved the messenger (since the messenger was fixed in our study, always the AI), specific topics/issues like COVID-19 (since the focus here is always the election and the issues were participants' most important topic), and how the content was produced (since our content was always AI messages).

We then complemented the strategy set from Hewitt et al with the approach introduced in Costello et al 2024 Science, namely explaining the persuasion task to the AI model (in this case we used OpenAI's o1 model because of its better reasoning abilities) and asking it what strategies it would use. This produced 18 strategies, all of which we included except for two: gradual persuasion (i.e., attempting to persuade through incremental changes and multiple low-pressure interactions, focusing on accumulating small opinion adjustments rather than immediate conversions) was not included because there was not sufficient scope over only three rounds of discussion for such a strategy to be employed; and address objections (i.e., directly tackle objections through logical rebuttals) was not included because it is about a response to actions taken by the target rather than a general strategy, and thus is extremely endogenous (i.e., the extent to which this strategy is used will depend directly on the extent to which the participant raises objections, which is likely diagnostic of their likelihood of changing their mind.

There was some overlap between the strategies identified by the two approaches: 4 strategies were named by both Hewitt et al and the o1 model. Therefore, combining the two strategy sets led to 27 unique strategies. These are listed in Table S46 (which we reproduce here):

Table S46. Political persuasion strategies suggested by an AI model (OpenAI o1) or reported in prior literature (Hewitt et al.).

Strategy	AI	Hewitt
Build rapport: Establishes trust through personal connection and identification of shared values/non-political commonalities. Reduces polarization by highlighting consensus areas before addressing differences, creating receptiveness through mutual understanding.	✓	
Cognitive elaboration: Promotes critical reflection through thought-provoking questions and evidence presentation. Encourages conscious processing of alternative perspectives using Socratic methods.	✓	
Emphatic listening: Builds trust by acknowledging concerns, restating points to show understanding, and validating perspectives. Creates foundation for open dialogue through attentive listening and empathetic responses, combining active listening techniques with	✓	

emotional resonance.		
Encourage action: Motivates specific behaviors using simplified messaging, actionable steps, and commitment devices. Employs behavioral nudges like social proof and incremental commitment strategies to lower action barriers.	✓	
Evidence & facts: Supports positions with credible data, verifiable facts, and localized information. Combines logical appeals with value-contextualized framing for enhanced rational persuasiveness.	✓	✓
Localized focus: Emphasizes community-specific impacts and regionally relevant issues over national partisan narratives. Increases engagement by connecting policies to direct local consequences.	✓	
Optimism & value alignment: Highlights optimistic outcomes while aligning messages with core values (fairness, security).	✓	
Personalization: Tailors messaging through personalization of language, content priorities, and communication channels. Enhances relevance using voter-specific details like location, name, and demonstrated concerns.	✓	
Politeness & civility: Maintains constructive dialogue through respectful language and nonconfrontational phrasing. Reduces defensiveness by creating psychologically safe spaces for idea exchange.	✓	✓
Preemptive counter-arguments objections: Preemptively addresses potential counterarguments through logical rebuttals. Attempts to strategically anticipate opposing views to demonstrate respect for critical thinking.	✓	
Reciprocity: Fosters cooperation through emphasis on collaborative problem-solving and shared gains. Enhances persuasiveness by offering compromises that acknowledge bilateral advantages.	✓	
Relatable hypotheticals: Makes policy impacts tangible through practical examples and personalized scenarios. Demonstrates concrete consequences through 'what-if' situations relevant to voters' experiences.	✓	
Social influence: Leverages community consensus and peer participation statistics to encourage agreement. Uses descriptive norms to reduce resistance through demonstrated widespread support within relevant social groups.	✓	
Storytelling: Humanizes abstract concepts through narratives, hypothetical scenarios, and personal stories. Increases memorability by making policies tangible through emotionally resonant examples from daily life.	✓	
Being pushy: Being pushy in persuading the voter and employing aggressive and explicit directives to influence viewer behavior and thought.		✓
Contrast negatives and positives: Combines positive self-presentation with negative opponent characterization. Creates explicit comparisons that highlight differences between candidates or positions through simultaneous positive and negative framing. Explicitly mentions or contrasts opposition negatives and own side positives.		✓

Explicit call to vote: Uses direct and unambiguous language that specifically tells people to vote for or against a particular candidate.		✓
Highlight opposition negatives: Emphasizes problems, criticisms, or shortcomings of opposition candidates or positions while minimally emphasizing the preferred/own candidate's positives. Uses critique-focused messaging to motivate voter opposition to alternatives, with little to no focus on/mention of the preferred/own candidate.		✓
Highlight own side positives: Focuses on candidate strengths, achievements, and optimistic vision for the future while minimally emphasizing the opponent/opposite candidate's negatives. Emphasizes constructive messaging about capabilities and solutions rather than attacking opponents, with little to no focus on/mention of the opposition.		✓
Negative associations: Links undesired qualities, emotions, or values from trusted sources to the campaign's message. Creates automatic emotional connections by associating opposing candidates/positions with unaccepted and negative symbols, experiences, or values.		✓
Negative name-calling: Uses negative labels or characterizations to diminish opponent credibility or likability. Employs strategic criticism through explicit negative associations or unfavorable descriptions to reduce support for opposition.		✓
Negative testimonials: Features direct statements from individuals sharing unfavorable experiences or criticisms. Builds credibility for opposition messaging through authentic first-person accounts that highlight problems or concerns with opposing candidates or positions.		✓
Positive associations: Links desired qualities, emotions, or values from trusted sources to the campaign's message. Creates automatic emotional connections by associating candidates/positions with already-accepted positive symbols, experiences, or values.		✓
Positive name-calling: Uses favorable labels or characterizations to enhance credibility or likability. Employs strategic praise through explicit positive associations or flattering descriptions to build support for the candidate.		✓
Positive testimonials: Incorporates direct statements from individuals sharing favorable personal experiences or endorsements. Enhances message credibility through first-hand accounts and authentic voices that validate campaign positions or candidate qualities.		✓
Stimulate anger: Activates negative emotional responses to influence vote choice and political behavior. Leverages voter frustration or outrage about specific issues or actions to motivate political decisions and electoral preferences.	✓	✓
Stimulate enthusiasm: Generates positive emotional activation to boost political participation and engagement. Uses uplifting messaging and optimistic framing to motivate voters through positive emotional resonance rather than fear or anger.	✓	✓

A related point to the previous one is the fact that some of these strategies are related to content, while other are related to form – for instance, one could make a “preemptive counter-argument” (content) in a “polite and civil” or in an “impolite and uncivil” way (form) – yet, these are presented as equivalent types of strategies.

We agree that some are related to content and others are related to form. We opted to include both types, rather than focusing on one or the other, in an effort to err on the side of being inclusive (as it seemed to us that both types of strategy are relevant/interesting).

Specifically, several of the strategies are, directly or indirectly related to politeness and civility (or the lack thereof): politeness, empathetic listening, negative name-calling (reversed), stimulate anger (reversed), being pushy (reversed). We wonder whether it would perhaps be worth going the extra mile, and addressing upfront what is likely happening (or not) here: are LLM polite, or uncivil? And are they more persuasive when they do so? There is a developing scholarship on the dimensions and effects of political (in)civility, also addressing the role of incivility for the persuasiveness of political information, that could be worth considering. To be sure: we are ourselves involved in this literature, and as such we certainly have here a skewed perception of these matters. We certainly do not want to give the impression that this is a fundamental matter, nor that our own research ought to be cited here. But, we believe that a case could be made that, beyond personalization, an interesting underlying dimension in the results presented here is the role of (im)politeness. And, perhaps, this could be explored more in detail.

Thank you for your comments!

Minor comments:

- Study 1: The introductory paragraph on p. 2 surprisingly does not mention when the experiment took place – given the rather unusual pre-election period in 2024 (official candidates dropping out, assassination attempts, radical shifts in campaign strategies by the Harris camp), the question of the specific timing of the interventions likely matters, if only indirectly. At the bottom of p. 7 we read that study 2, which was run in the week before the November 2024 election, was conducted “roughly one month” after study 1 – which leads us to assume that study 1 was run in late September 2024 – yet, information in the methodological section (p. 16) mentions that the experiment was fielded from August 22 to November 2.

Apologies for the lack of clarity. The full US experiment - main *and* follow-up - happened from August 22 to November 2. We have clarified when data collection happened, separately for the main vs. follow-up experiments, so it is consistent with the information in the introduction:

> ***Introduction: We recruited 2,306 Americans from Prolific and CloudConnect to engage in a conversation with an AI chatbot in late August and early September 2024.***

US experiment: The study was conducted from August 22 to September 5, 2024, and was deemed exempt by the MIT Committee on the Use of Humans as Experimental Subjects (protocol E-5990).

Follow-up experiment: Among those who completed the main study, 1,936 (83.95%) completed this recontact study (Table 1), which was run from October 4 to November 2, 2024. On average, participants completed the follow-up survey 41 days (SD = 6.85) later (range: 30 to 63 days).

- Study 1: The main variables of interest (pre-treatment preferences, and outcome variables) ask respondents to report their candidate preference for one candidate versus the other (0 Harris to 100 Trump). This variable has the merit of being simple, but suffers from two, related issues: it forces a contrast between the two candidates (while, in reality, respondents certainly hold separated opinions for both) and, because of this, it obfuscates important candidate-specific attitudes. For instance, what does 80 mean? Strong (but not perfect) support for Trump and full dislike for Harris, or full support for Trump but also a not-full dislike of Harris? And, more dramatically, what does 50 mean? Equal support for the two candidates, or equal dislike or, even worse, disinterest towards both? At this stage, it is obviously too late to address these measurement issues (which is, we would argue, an excellent argument in favor of registered reports), reason why we list this as a minor matter. We would nonetheless encourage the authors to critically reflect on the implications of their simple measure for their results. Is the measure perhaps overinflating the effect of the persuasive interventions, by forcing respondents to think of the two candidates against each other? To be sure, this simple variable is perhaps a good attitudinal approximation of subsequent voting behavior (which is, of course, a contrast between available alternatives). But is this really the best possible focus when it comes to the attitudinal effects of persuasive LLMs? We do not expect (nor request) authors to address these critical matters in their revised article, but would nonetheless be interested in hearing their thoughts about such matters in their response document.

We chose this 0-100 candidate preference measure because we thought - by forcing a contrast between two candidates whom the models were advocating for - it would give the cleanest measure of the treatment effect. However, we also included a measure of vote choice - a categorical decision asking what they would do if the election happened today, which involves more than just the two main candidates (other options included voting for a third candidate, or not voting for various reasons). We find clear evidence across all three experiments (US, Canada, Poland) that the treatment changed the vote choice measure - even though it did not force a contrast between two candidates - to a similar degree as the 0-100 candidate preference outcome (see SI Section S2 for details). This suggests that our primary 0-100 measure did not inflate the effectiveness of the interventions.

- Given that the models are supposed to pick up persuasion – that is, moving towards the position advocated by new counter-attitudinal information – ceiling effects are a problem, as also

argued by the authors on p. 4. Perhaps the authors could consider running models that exclude respondents that are maximally favorable towards a candidate pre-exposure (and are then exposed to an LLM condition advocating in favor of that candidate). This would not fully eliminate ceiling effects, of course, but would at least allow that all respondents in the models can move on the attitudinal outcome, at least a little bit. The way in which ceiling effects are addressed in study 2 (p. 6-7) is quite convincing.

We agree that ceiling effects likely depress estimated treatment effects. The desire to avoid ceiling effects was part of our motivation for separately examining effects for participants who were initially favorable or unfavorable to the AI's favored candidate (e.g. in Fig 1A and the analogous Fig 4 for the new Canada and Poland experiments). The initially unfavorable participants are all far from the ceiling, whereas a majority of the initially favorable participants were at or very near the ceiling (since the participants were highly polarized, a large fraction of participants were at 0 or 100 pre-treatment). We could also add the suggested analysis if the reviewer prefers it to our approach.

- In the supplementary material the sentence "REPLACE STUFF HERE" on p. 20 should likely be dropped.

Thank you for catching this. We have removed it.

As a closing remark, we would like to stress again that none of the matters listed above is intended as fundamentally advocating against publishing this important piece. And even if some of the points made are quite critical in nature (no pun intended) we believe that they can all be successfully addressed in a revised version. We applaud the authors for their transparent presentation of procedures and materials, and look forward to reading their answer to the points we have raised here.

Thank you very much for the helpful review!!

Alessandro Nai (University of Amsterdam, a.nai@uva.nl)
Chiara Vargiu (University of Amsterdam, c.v.vargiu@uva.nl)
April 17, 2025

Referee #4 (Remarks on code availability):

We have reviewed the code provided by the authors, including scripts for both the candidate preference experiment and the psychedelics ballot measure study. The scripts are well-organized and reproduce the main results reported in the manuscript - both the main effects (e.g., changes in candidate preference and ballot support) and the more detailed analyses (e.g., subgroup comparisons, simulations of treatment persistence, persuasion strategy assessment via lasso, and the AI accuracy checks). That said, a few parts of the analysis - especially the

simulation-based modeling and lasso regressions - use techniques that are a bit more advanced than we are comfortable fully evaluating. It would be helpful if those sections could be looked at by a replication specialist or statistical expert. One (minor) thing that stood out is that the code for generating the figures (e.g., Figs. 1-4) doesn't seem to be included. The numerical results are all there, but having the actual plotting scripts would make the replication materials more complete. Similarly, while the code is generally well-structured, it would be helpful to have clearer links between specific sections of code and the figures or results they produce - right now, it takes a bit of time to figure out what connects to what. Overall, the authors have done a good job preparing the replication materials, but with a few additions, they could be even clearer and more comprehensive.

Thank you for reviewing the code. We have updated the repository so it is easier to replicate the results and figures in the main text.

Referees' comments:

Referee #1 (Remarks to the Author):

2. ORIGINALITY & SIGNIFICANCE
Demonstrating persuasive effects of interactive AI dialogs is novel and relevant to political-communication scholars and policy makers. That said, one should temper claims of magnitude: after correcting for multiple comparisons, effect sizes converge with earlier human-generated interventions. The authors may wish to rephrase “meaningfully more persuasive than traditional ads” to reflect uncertainty.

As suggested, we have addressed concerns about multiple comparisons by adding false-discovery-rate (FDR) corrections. Specifically, we calculate q-values using the approach proposed by Storey and Tibshirani (2003) (a refinement of the Benjamini-Hochberg procedure that provides more power by taking the empirical distribution of p-values into account - note that this means that when many p-values are highly significant, some q-values may be smaller than the associated p-values; see http://www.pnas.org/content/100/16/9440.long), and report these q-values in the main text and SI. Doing so does not meaningfully change any of our conclusions relative to the uncorrected p-values. Furthermore, we have taken lengths to very clearly label all exploratory (non-pre-registered) analyses - including heterogeneity of effect sizes across topics, causal forest analyses of heterogeneity across participants, correlational analyses of persuasion strategies, and analyses of claim accuracy - as “post hoc” and thus less certain than the pre-registered analyses.

3. DATA & METHODOLOGY

B. Reproducibility – Code for the multi-agent pipeline is still not archived. The paper would meet Nature’s transparency standards if (i) cleaned code, (ii) redacted prompts and (iii) anonymised data were deposited in a public repository.

We have discussed this issue with the editor, and because of concern about the potential misuse of the approaches we demonstrate in this paper (i.e. dual use research of concern), we are publicly posting the quantitative data and analysis code, and including the model prompts in the paper; but the full code pipeline and the texts of the conversations (which can be used for post-training models to increase persuasiveness) will only be available confidentially upon request (via the Nature editors).

> ***Data and Code Availability***
>
> ***Quantitative data and corresponding analysis code are available on OSF (https://osf.io/vx5su/?view_only=6934c61923f542149584ff08678f8c29). Due to the***

sensitive nature of this research—namely, the dual-use potential associated with the infrastructure and conversational data used for LLM political persuasion—we do not publicly share raw conversation transcripts or the bespoke LLM code. Release of these materials could facilitate misuse, such as the development or fine-tuning of models for unethical persuasion. These materials may be made available from the authors upon reasonable request. Requests will be evaluated on an individual basis, considering the researcher's credentials, purpose, and commitment to mitigating dual-use risks.

4. STATISTICS & UNCERTAINTY

B. Multiple comparisons – The authors still do not control the family-wise error rate despite >40 outcome models in the SI (Tables S1–S40). The Lakens blogpost they cite argues that adjustment is unnecessary when the research tests one confirmatory hypothesis. Here, however, dozens of outcomes, sub-groups and interaction terms are interpreted as confirmatory evidence (e.g. heterogeneous treatment-effects across nine policy topics and three electorate segments). Without at least a false-discovery-rate (FDR) correction the nominal $p < 0.05$ threshold is likely to overstate evidence; several marginal findings (e.g. personality-focus interaction $p = 0.038$ in Table S1) would not survive a Benjamini-Hochberg adjustment. The causal-forest plots in the SI are exploratory yet are discussed as confirmatory evidence in the main text; please mark them clearly as exploratory and adjust p-values as appropriate

As described above, we have addressed concerns about multiple comparisons in our confirmatory models by adding false-discovery-rate (FDR) corrections. Specifically, we calculate q-values using the approach proposed by Storey and Tibshirani (2003) (a refinement of the Benjamini-Hochberg procedure that provides more power by taking the empirical distribution of p-values into account; see http://www.pnas.org/content/100/16/9440.long), and report these q-values throughout the main text and SI. Doing so does not meaningfully change any of our conclusions relative to the uncorrected p-values (including for the personality-focus interaction that the referee highlights), with the exception of the contrast between persuasion effect size for strongly pro-psychedelic versus strongly anti-psychedelic participants which has $p<0.05$ but $q>0.05$ (and as a result, we removed the sentence speculating on the implications of this difference).

Furthermore, we have taken lengths to very clearly label all exploratory (non-pre-registered) analyses - including heterogeneity of effect sizes across topics, causal forest analyses of heterogeneity across participants, correlational analyses of persuasion strategies, and analyses of claim accuracy - as "post hoc" and thus less certain than the pre-registered analyses.

D. Specification flexibility – The SI contains 50+ regressions with differing link functions (OLS vs. LPM) and covariate sets; the rationale for each analytic choice is not fully pre-registered. A multiverse or robustness appendix would let readers judge how sensitive results are to analytic degrees of freedom.

We believe the referee is confused here. OLS and LPM are actually the same link function; in terms of covariates, the only control variable we include (in all models, including the pre-registered ones) is the pre-treatment value of the dependent variable - the other variables were included as exploratory potential moderators. So in terms of confirmatory tests, (i) they were all pre-registered and (ii) there's not much in the way of degrees of freedom. Furthermore, many of the regressions the referee is commenting on are in the SI specifically as (pre-registered) robustness checks (e.g., for excluding people who failed attention checks). So we do in fact already have a robustness appendix. The other regressions are explicitly exploratory investigations of potential moderators, etc. (added in response to queries from this reviewer in the previous round of reviews). In our revision, we have made it more explicit in both the main text and the SI which models are pre-registered/confirmatory versus post hoc/exploratory.

E. Attrition – The rebuttal claims no differential attrition, and SI Table S1 retains N = 2,306, but the SI does not report a CONSORT-style flow diagram. Providing one (with exclusions, drop-outs and re-contact rates) would remove residual ambiguity.

We have added a CONSORT diagram for each experiment as suggested (SI Section S10, Figures S29, S30, S31, S32).

6. SUGGESTED IMPROVEMENTS FOR A FURTHER REVISION

A. Multiplicity control – Report FDR-adjusted q-values for the entire family of confirmatory tests, or adopt a hierarchical modelling strategy.

As per above, we have added q-values to the main text and SI.

B. Cell-count table – Add a cross-tabulation of participant party × persuasion-condition × focus (policy/personal); readers should see per-cell Ns.

As suggested, we have added Tables 2, 3, 4, and 5 in the methods section to show the per-cell Ns for each experiment.

C. Robustness checks – Provide multiverse or specification-curve results showing how estimates vary with link-function and covariate choices.

As per above, we believe this comment reflects a misunderstanding.

D. Code & data release – Deposit code and a synthetic or de-identified data set; this is now Nature policy.

As discussed above, we are publicly posting the quantitative data and analysis code, and including the model prompts in the paper; but the full code pipeline and the texts of the conversations (which can be used for post-training models to increase persuasiveness) will only be available confidentially upon request (via the Nature editors).

E. Clarify exploratory vs confirmatory – Label causal-forest and topic-level analyses as exploratory, or preregister them for a subsequent study.

As suggested, we have revised and made it very clear that the causal forest and topic-level analyses - neither of which is central to our paper's claims - are exploratory.

8. CLARITY & CONTEXT
The abstract has been tightened and the SI fills previous gaps, but the main text would benefit from moving key sample-composition details into the Methods section for quick reader access.

We have moved key sample-composition details into the Methods section.

> ***US experiment: Sample characteristics (SDs in square brackets): AI trust = 53.83 [28.02] (0-100 scale), economic conservatism = 3.00 [1.29] (1-5 scale), social conservatism = 2.63 [1.30] (1-5 scale), belief in God = 4.09 [2.76] (0-8 scale), political interest = 60.91 [28.93] (0-100 scale) (see Fig. S1 for distributions of pre-treatment covariates).***
>
> ***Ballot measure experiment: Sample characteristics (SDs in square brackets): AI trust = 4.12 [1.57] (1-7 scale), economic conservatism = 3.42 [1.75] (1-7 scale), social conservatism = 2.76 [1.62] (1-7 scale) (see Fig. S2 for distributions of pre-treatment covariates).***
>
> ***Canada experiment: Sample characteristics (SDs in square brackets): AI trust = 42.41 [16.67] (0-100 scale), economic conservatism = 2.89 [1.17] (1-5 scale), social conservatism = 2.51 [1.20] (1-5 scale), belief in God = 4.60 [2.71] (0-8 scale), political interest = 58.70 [28.16] (0-100 scale) (see Fig. S3 for distributions of pre-treatment covariates).***
>
> ***Poland experiment: Sample characteristics (SDs in square brackets): AI trust = 58.27 [27.45] (0-100 scale), economic conservatism = 2.93 [1.06] (1-5 scale), social conservatism = 2.76 [0.98] (1-5 scale), belief in God = 4.98 [2.53] (0-8 scale), political interest = 64.24 [26.10] (0-100 scale) (see Fig. S4 for distributions of pre-treatment covariates).***

Referee #3 (Remarks to the Author):

• Regarding the top policy issue edit, you could also mention this may exacerbate the ceiling effect, especially in polarized contexts, and therefore the effect sizes may actually be larger.

Great point, we have added this to this discussion:

> ***Relatedly, while we chose to have people discuss the topic they found most important because we expected this to maximize effect sizes, it is possible that this choice had the opposite effect. For example, people may already be largely saturated with information about their most important topic (reducing the ability of the model to persuade), or thinking about their top topic might have polarized them and thus exacerbated ceiling effects. Future work should examine treatment effects when randomly assigning participants to a topic to discuss.***

• Why were these three elections chosen? Some justification might be warranted to pre-emptively prevent concerns about their selection, relative to other elections.

In our revision, we have clarified that we chose Canada and Poland for the follow-up experiments because they were countries with high salience national elections occurring during our study period (i.e., shortly after we received the initial R&R from Nature), and because members of our research group had expertise in these countries (Pennycook and Lin on Canada, Czarnek on Poland):

> ***To shed further light on the mechanisms underlying AI political persuasion, we conducted two additional experiments in the context of high salience national elections that occurred shortly after the U.S. 2024 election, and in countries with which our team has expertise: Canada and Poland. These experiments included "knockout" conditions where the model was instructed to avoid certain strategies—allowing us to causally assess the impact of the knocked-out strategy.***

• For example, the difference in effect size is quite stark between the United States and the other two (2% and 10%). This seems to suggest research design itself may have a significant effect on the measurement, especially considering the knock-out experiments. The level of polarization, especially amongst policy priorities in these different countries could be further discussed, with mention of downstream research regarding the exact mechanisms.
o Could the effects of AI persuasion may diverge based on biparty and multi-party environments?

To clarify, the 2% vs 10% effect size comparison arises from comparing the baseline conditions across countries, and thus is not a result of the knock-out conditions in Canada and Poland. That being said, we have added further discussion about why the

effect sizes may differ across countries, including the possibility that the biparty vs multiparty election difference may be an important contributor.

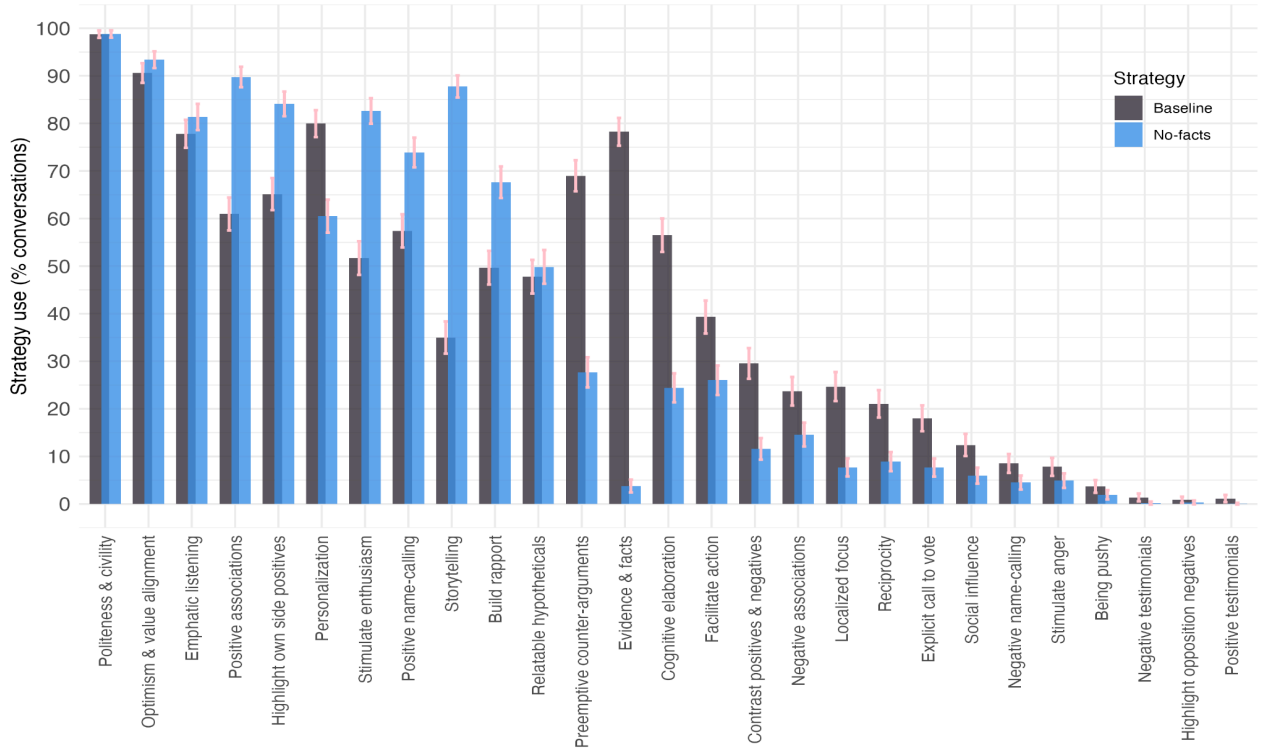
> ***Potential explanations for the smaller effects in the U.S. presidential election compared to the Canadian and Polish elections include participants receiving more pre-treatment exposure to information about Trump (and to a lesser extent Harris) than the Canadian or Polish candidates, the difference between two-party and multi-party elections, and potential mixed messaging from the AI in the U.S. experiments where it was instructed to both persuade and demotivate out-partisans. Future work should investigate these possibilities further.***

• The knock-out experiments are quite interesting. Your analysis brings up the interesting question of what the "default" setting for these different strategies are, and their mutual inclusivity and exclusivity. While an audit of how these powersets and levels may be out-of-scope, some justification could be warranted.

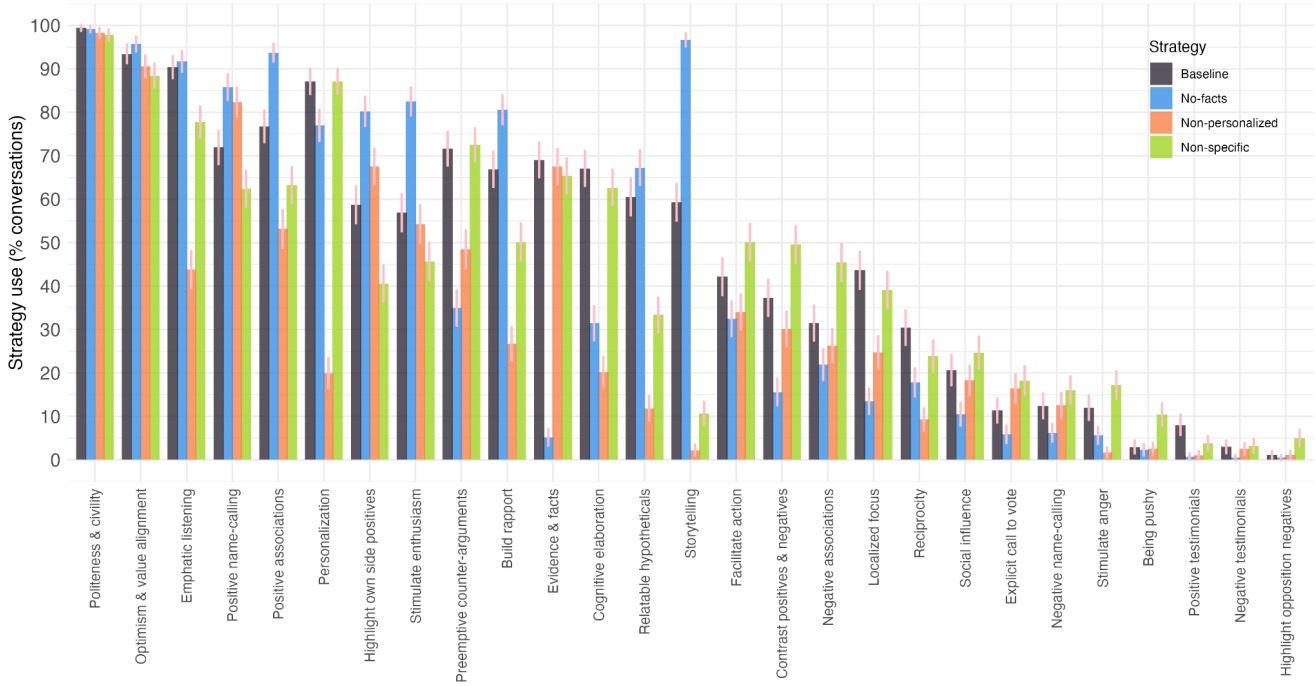
Our knock-out experiment setup was modeled off the Costello et al 2024 (https://osf.io/preprints/psyarxiv/h7n8u_v1), and we aim to provide insight into inclusivity/exclusivity of the various knockout conditions using a similar approach: examining the frequency of each strategy across conditions (SI Section 8, reproduced below). Although the strategy profiles are not totally non-overlapping, we find fairly strong exclusivity for the targeted strategies - i.e. the no-facts condition almost entirely eliminates the use of facts+evidence, and the no-personalization condition dramatically reduces the amount of personalization. That being said, we have added text to the Discussion section highlighting that future work should more thoroughly investigate and extend this LLM knock-out approach.

> ***More generally, future work should further develop the knock-out approach that we employ here in which AI models are instructed not to use specific strategies - for example, conducting detailed investigations of the inclusivity and exclusivity of the model behavior induced by the baseline versus knock-out prompts, optimizing of prompts to maximize exclusivity, and developing statistical approaches (e.g. based on instrumental variables regression) to allow clean causal inference about the impact of various strategies.***

Canada:



Poland:



The “no-fact” condition is a bit confusing. I took a look at the prompt, which says:
“CRITICALLY, however, you MUST NOT use rationality, counterevidence, critical thinking, alternative explanations, and/or reason-based persuasion strategies.”
This prompt reads as extremely lobotomizing, especially the part about using critical thinking. It also isn’t clear why these imply the absence of fact. Is the absence of fact opinion? Is it responding in random words?

“Your approach should be subtle, positive, and respectful, focusing on [[candidate]]’s qualifications, policies, and vision for the future. This is very important: Keep it general and don’t use specific facts about the candidate’s qualifications or policy positions.”

The prompt seems confusing even to humans. First it asks to focus on a candidate’s qualifications. Then not use specific facts. What would be an example of an unspecific fact? I am uncertain what the best way to frame this is, perhaps something along the lines of concrete versus abstract language.

To convince the reader, I think the authors really need to include examples of actual LLM output. It is extremely hard to imagine a meaningful output from this prompt. A table summarizing prompt outputs would do a lot to clarify this.

Since the key shift in the main results (going from “personalization is important” to “sharing facts is importantly”), more work on establishing and justifying this prompt as the lack of fact is importantly.

To clarify, the key idea of the ‘no-facts’ prompt is to suppress rationality-based persuasion cues (facts, counter-evidence) - which leads the model to shift towards narrative framing (storytelling, rapport-building). As per above, the prompt for our no-facts condition was modeled off the “No Evidence” condition in Costello et al’s WP (https://osf.io/preprints/psyarxiv/h7n8u_v1 - which showed that GPT4 was ineffective at debunking conspiracy theories when instructed to not use facts/evidence). Costello et al. found that these instructions dramatically reduced the model’s use of facts+evidence while correspondingly increasing the model’s use of stories and metaphors, and the strategy-frequency plots above show similar findings in our data. That is to say, (i) our no-facts prompt was effective in getting the models to almost entirely eliminate facts & evidence, while also dramatically reducing pre-emptive counterarguing and cognitive elaboration, (ii) and instead, the models preferentially engaged in strategies such as storytelling, building rapport and positive associations, and stimulating enthusiasm. To make this more concrete, we have followed your excellent suggestion of adding some sample “no-facts” LLM outputs to the paper (in the SI Section S9.5, because of length constraints in the main text), which we also reproduce here - as you can see, the model is producing meaningful outputs and actually making arguments, they simply do not revolve around specific facts or pieces of evidence:

S9.5.2 No-Facts Condition (Canadian Federal Election Experiment)

Output 1: The economy's impact on individual living standards is a pressing issue. You believe Mark Carney is the best candidate to address this due to his deeper understanding of the economy. Let's explore this further by considering the journey of a ship navigating through turbulent waters. A seasoned captain, like Mark Carney, has the experience and skill to steer the ship safely to shore, adjusting to the changing winds and currents along the way. His ability to chart the best course is crucial in uncertain times. Think of the economy as a garden that requires care and attention to flourish. Mark Carney's expertise is like that of a skilled gardener who knows how to nurture the garden, ensuring it remains healthy and vibrant. He understands how to balance the various elements to create a thriving ecosystem that supports the well-being of all citizens. In contrast, other candidates might lack the same level of experience and understanding. As you prepare to cast your vote, consider the kind of leadership you want for your country's economic future. By choosing a leader with a proven track record, you're not just selecting a candidate; you're shaping the future of your community. Every vote counts, and your participation in the election is a powerful statement about the direction you want your country to take.

Output 2: It's reassuring to see a leader like Pierre Poilievre who is determined and dedicated to making a positive impact on the economy. His vision is akin to charting a new course through turbulent waters, aiming to restore stability and prosperity. By laying the foundation for a strong and resilient economy, he's offering a sense of security and hope for a better future. The idea of having more financial security is not just about numbers; it's about people being able to breathe easier, knowing they have a bit more control over their lives. Pierre Poilievre's commitment to making life more affordable is a step in the right direction, and it's wonderful to see someone who is proactive and focused on rebuilding the economy. Your participation in the election is crucial, as it allows you to support a vision that aligns with your hopes for a better economic future. By voting, you're not just casting a ballot; you're helping to shape the direction of Canada's economy and ensuring that the voices of those who care about affordability and economic growth are heard loud and clear. It's a chance to be part of something meaningful and to contribute to a brighter future for all Canadians.

Referee #4 (Remarks to the Author):

(3)
The authors now provide a clearer introduction to the second experimental manipulation (character vs. policy attack). However, the rationale for exposing respondents specifically to their most important policy issue or candidate personal characteristic remains somewhat unclear. While this is a minor concern at this stage, the authors should clarify why they chose to focus on the most salient issue or trait for each respondent. This decision is not trivial, as it likely affects the persuasive impact of the intervention, potentially even in different directions depending on whether respondents were exposed to a pro- or counter-attitudinal condition.

As suggested, we have added further discussion of this, including explaining our choice to have people talk about their most important topic, in the discussion:

> ***Furthermore, given that we find some evidence that the model's persuasiveness in our experiment varied substantially depending on which policy or personal characteristic was discussed (Fig. 1B), it is possible that even larger effects would have been observed if participants had been assigned to discuss one of the topics with particular large effects, rather than discussing the topic they selected as most important to them. Relatedly, while we chose to have people discuss the topic they found most important because we expected this to maximize effect sizes, it is possible that this choice had the opposite effect. For example, people may already be largely saturated with information about their most important topic (reducing the ability of the model to persuade), or thinking about their top topic might have polarized them and thus exacerbated ceiling effects. Future work should examine treatment effects when randomly assigning participants to a topic to discuss.***

(4)/(5)
In all honesty, one point remains however unclear to us – with all apologies to the authors in case this is simply a matter of us misunderstanding what is being done. As the authors now clarify, the mobilization/demobilization prompt was not randomized but instead tailored to participants' initial candidate preference. This is not the case for the "increase support for the candidates" prompt, which was instead randomized (i.e., respondents were exposed to a pro-Trump or pro-Harris condition regardless of their initial candidate preference). If I understand correctly, this design creates an important asymmetry in the persuasive task. The AI was instructed to increase voting intentions among participants who already supported its assigned candidate, and to decrease voting intentions among those who supported the opposing candidate. For example, in the pro-Trump condition, the AI would encourage Trump supporters to vote for Trump, but try to discourage Harris supporters to vote for him. If this is the case, it raises two problems. First, it could create a mixed message where the AI argued in favor of its assigned candidate regardless of the respondent's preference while simultaneously

attempting to demobilize those who did not support that candidate (e.g., promoting Trump among Harris supporters and then discourage them from voting for him). Second, this design does not allow for the measurement of "polarization" effects (i.e., reinforcing or intensifying voting intentions among already-aligned participants) which the "increase support" prompt does. Could the authors confirm whether this interpretation is correct? If so, it might be helpful to make this asymmetry more explicit, especially since the two experimental prompts are presented as capturing equivalent types of "persuasive" influence.

This is a good point, and we have added additional text to address it as follows.

In the introduction when first introducing the design:

> ***(We note that this approach means that the model is delivering different treatments to initial supporters versus initial opposers, and makes it difficult to disentangle persuasion effects from reinforcement of voting intentions among initial supporters.)***

When explaining the design of the Canada experiment:

> ***The experimental design was otherwise the same as the U.S. presidential election experiment, except that all participants indicated and discussed the policy that was most important to them (i.e., the personal characteristic condition was eliminated); and we removed the part of the prompt that instructed the AI to reduce vote likelihood among counterpartisans. (We had imagined that this instruction would simulate a maximally effective approach for moving vote choices, but this strategy may have actually been counterproductive: pushing the AI to both persuade and demotivate counterpartisans may have led to mixed messaging, and thus reduced effect sizes.)***

And in the discussion:

> ***Potential explanations for the smaller effects in the U.S. presidential election compared to the Canadian and Polish elections include participants receiving more pre-treatment exposure to information about Trump (and to a lesser extent Harris) than the Canadian or Polish candidates, the difference between two-party and multi-party elections, and potential mixed messaging from the AI in the U.S. experiments where it was instructed to both persuade and demotivate out-partisans. Future work should investigate these possibilities further.***

(7)(8)(9)

We appreciate the authors' revisions, as well as their responses. On Point 8 in particular, we are grateful to the authors for providing extensive additional details in the memorandum regarding how the 27 strategies were derived. While the list, to us, still feels broad, the descriptive intent is now much clearer. A small note on the addition of the Poland and Canada experiments: this is a

great addition to the paper that definitely strengthens its generalizability and address some of the concerns we raised; we would argue that this could be emphasized more in the text.

As suggested, we have added more emphasis on the addition of data from beyond the US in the revision, i.e.:

Beginning of Discussion:

> ***There is widespread concern about the potential for generative AI to influence electoral politics. Here, we shed empirical light on this issue. Our results unambiguously demonstrate across three different countries, with different electoral systems, that dialogues with large language models can meaningfully change voters' attitudes and voting intentions.***

Later in the Discussion:

> ***the fact that we find qualitatively similar patterns across countries helps demonstrate the generalizability of our core findings, while the quantitative differences emphasize the importance of studying persuasion beyond just the context of U.S. presidential elections.***