Dialogues on Democracy: Belief-Tailored AI Conversations Reduce Inaccurate Election Denial Beliefs

Shaye-Ann Hopkins

smm197@duke.edu

Vienna University of Economics and Business

Thomas Costello

Carnegie Mellon University

Gordon Pennycook

Cornell University https://orcid.org/0000-0003-1344-6143

David Rand

Massachusetts Institute of Technology https://orcid.org/0000-0001-8975-2783

Article

Keywords: AI, LLM, elections, conspiracies, belief updating, election fraud

Posted Date: January 28th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-8663921/v1

Additional Declarations: There is **NO** Competing Interest.

Dialogues on Democracy: Belief-Tailored AI Conversations Reduce Inaccurate Election Denial Beliefs

Shaye-Ann Hopkins* [1,2], Thomas Costello [3,4], Gordon Pennycook [5], & David Rand [4,5]

[1] Duke University

[2] Vienna University of Economics & Business

[3] Carnegie Mellon University

[4] Massachusetts Institute of Technology

[5] Cornell University

*__Corresponding author__: Shaye-Ann Hopkins; shaye-ann.hopkins@wu.ac.at

Abstract

Elections, while central to democratic functioning, have become increasingly threatened by beliefs about election fraud. Artificial intelligence (AI) provides a unique opportunity to address such false beliefs through dynamic conversation and debunking. Across two experiments conducted shortly before the 2024 US Presidential election (N = 1,768 Republicans from Lucid who endorsed election fraud claims), we examined whether corrective conversations with a large language model can reduce election denial beliefs through claim-targeted dialogues. We tested the effects of two treatments (an information-tailored and values-tailored AI dialogue) in which AI fact-checked their claims and tailored arguments to the specific election conspiracy that the participants themselves articulated. Both treatment conditions, when compared to a control dialogue or a simple statement that the conspiracy was incorrect, reduced confidence in their election conspiracy claims and more general beliefs of 2020 election fraud. There was no significant difference between information-tailored and values-tailored feedback. Promisingly, participants with the strongest baseline denialism experienced the largest decreases in denial beliefs. These studies highlight the potential of AI-driven interventions to address election misinformation.

Keywords: AI, LLM, elections, conspiracies, belief updating, election fraud

Dialogues on Democracy: AI Conversations Effectively Correct Inaccurate Election Denial Beliefs

Confidence in electoral systems and associated actors is vital to a well-functioning society. However, recent election cycles, particularly within the US, have been increasingly threatened by issues such as decreasing trust, election misinformation, and conspiratorial narratives [1,2]. For example, the US reached an all-time low in public trust in the government in 2019, where only 17% of the public reported that they trusted their government [3]. Furthermore, echo chambers and echo platforms, along with selective content exposure, have been linked to polarized media environments [4] and more prevalent inaccurate beliefs [5]. The 2025 Global Risk Report identifies mis- and disinformation and armed conflict as some of the most important short-term global risks [6], highlighting potential threats to societal cohesion, eroding trust, and exacerbated divisions within nations. Other risks associated with these beliefs include democratic backsliding and support for more authoritarian regimes [7].

In the current work, we focus on misconceptions about (lack of) election integrity in the United States, particularly among Republicans. Republican confidence in U.S. elections has been depressed for more than a decade [8], and several indicators point toward an ongoing downward trajectory [9,10]. Reactions to the 2020 presidential election further solidified this issue. A national post-election survey reported that 38% of Americans doubted that the outcome was fair, including nearly two-thirds of Republicans but only 11% of Democrats [11]. Much of this partisanship has been traced to repeated allegations of widespread fraud made by Donald Trump and other political elites [12]. Conversely, comprehensive audits of the 2020 election uncovered no meaningful irregularities [13], and studies of earlier U.S. elections similarly find little evidence of systemic fraud [14,15].

Despite the disparity between public beliefs and the empirical record, empirical attempts to counter false narratives about election fraud, especially among those most committed to such beliefs, have produced limited or null effects [16,17]. For example, Bailard et al. [18] found that fact-checks of Trump's 2020 fraud claims did not meaningfully raise Republicans' confidence in the election. Meanwhile, [19] showed that while specific falsehoods about the 2022 Arizona gubernatorial race could be corrected, such corrections failed to shift broader perceptions about election fraud or confidence in recent elections. This lack of meaningful corrective effects is in line with longstanding theories in political psychology.

Considerable research in psychology argues that people rarely revise their views in response to counter-attitudinal information, especially on identity-charged topics. Theories of motivated reasoning highlight that individuals process political evidence through the lens of partisan identity, cultural values, and prior beliefs, often downweighing or dismissing information that contradicts their worldview [20–23]. This makes these beliefs difficult targets for traditional debunking methods, which typically rely on presenting factual counterarguments. In some accounts, corrective information may even backfire, strengthening the very beliefs it aims to undermine [16,24].

In contrast to this perspective, however, recent research has found that conversations with a large language model can substantially change politically relevant attitudes. For example, AI dialogues can effectively debunk conspiratorial beliefs [25], and this primarily works by leveraging facts [26]. Similar effects have been observed in the context of attitudes towards presidential candidates [27] and policy stances [28,29]. Other research[30] has found that messages generated by GPT-4 were broadly persuasive, but micro-targeted messages were not statistically different from non-micro-targeted messages, even when manipulating the type and number of attributes used to tailor the message.

Overall, despite substantial progress in this field, existing research leaves open key gaps. There is an opportunity to explore 1) how to effectively correct beliefs for those with more extreme beliefs in more identity-

charged topics like elections and their legitimacy, and 2) if more hyper-personalized approaches (e.g., tailoring feedback to an individual's values, instead of demographic attributes) have added benefit. Given the ever-evolving nature of politics and a global shift towards authoritarianism in the past decade, there is benefit in understanding how evolving technology can be applied specifically within the context of election denialism.

Therefore, in the current work, we ask whether AI dialogues can effectively reduce misconceptions about election integrity among Republicans immediately prior to the 2024 presidential election. We examine effects both on attitudes specifically about fraud in the 2020 election, as well as belief in the participants' own unique beliefs (as articulated in a free-text response). We also examine whether allowing the model to target its persuasion to the participants' values improves persuasiveness (Study 1), and contrast treatment effects with both an "adversarial statement" that simply indicates that the belief is incorrect without providing evidence (Studies 1 and 2) and a pure control in which the AI talks to participants about dogs vs cats (Study 2). Our pre-registration, data, surveys, and code can be found on OSF (Study 1: https://osf.io/rq3fm, preregistered on Jul 29, 2024; Study 2: https://osf.io/pgdzw, preregistered on Sep 25, 2024). An overview of all conversations included in the analysis can be found here: https://shaye-hopkins.shinyapps.io/ai-election-convos/.

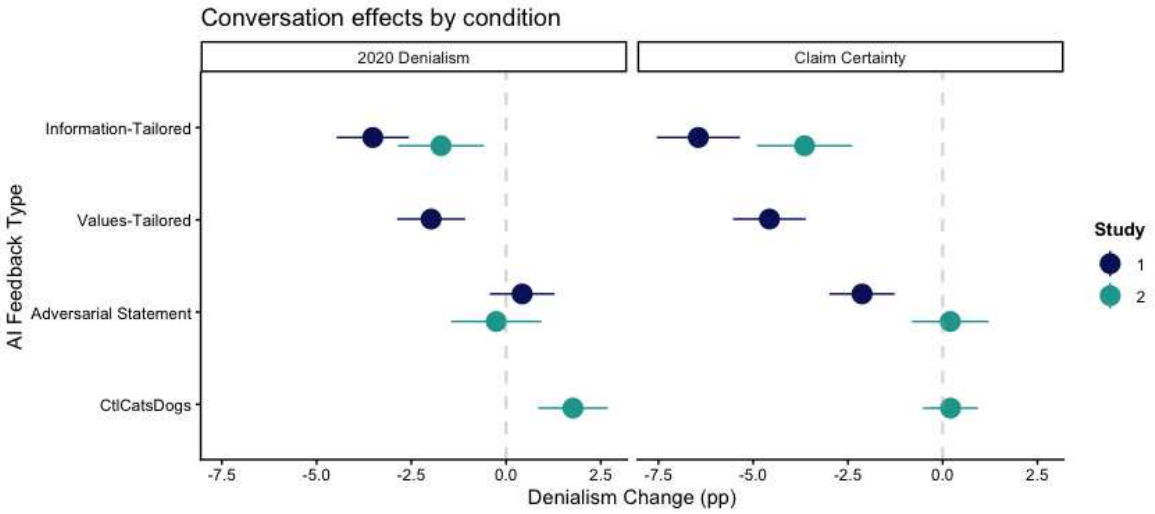
Results

We used ordinary least squares (OLS) linear regressions, predicting post-experiment election denial beliefs (2020 Election Denialism, Claim Certainty) using a treatment condition variable and controlling for mean-centered pre-intervention beliefs and interface type (Qualtrics or chatbot; in Study 1 only). Since there were no significant interactions between treatment condition and interface type (Supplementary Table C1), we focus on treatment effects and control for interface type. For all regression analyses, *p*-values and confidence intervals were computed using robust standard errors, and an alpha level of .05. Analysis was conducted using R Statistical language (version 4.5.1; R Core Team, 2025).

Belief-tailored AI dialogues reduce election denialism

Across both studies, the information-tailored feedback condition was the most effective intervention in reducing denialism outcomes by 5-8%. Significance was retained after controlling for demographics and other covariates (see Fig. 1 for effects across both studies and Supplementary Fig. C2).

Fig. 1: Denialism changes by condition & study.



Study 1. The primary outcome of interest was 2020 Denialism, or the belief that fraud changed the outcome of the 2020 election. For Study 1, relative to the adversarial statement condition, the information-tailored treatment (adopted from Costello et al. [25]) significantly decreased denialism beliefs around the 2020 election by 4 percentage points (pp) on a 0-100 scale, or 8% (b = -3.73, β = -.12, p = .003, d = -.20, Δ% = -7.69%). Values-tailored feedback also significantly reduced 2020 denialism by 2 points, or 5%, relative to the adversarial statement condition (b = -2.41, β = -.08, p = .042, d = -.16, Δ% = -4.98%). Estimated marginal means indicated that the difference between the two treatment conditions was not statistically different (values vs. information: p = .560).

For Claim Certainty (or certainty about the specific election fraud claim they made), we saw slightly stronger effects. The information-tailored feedback condition again decreased person-specific beliefs by over 4 pp, or 5% (b = -4.12, β = -.21, p = .002, d = -.30, Δ% = -5.07%), while values-tailored feedback did not (p = 0.081). Again, information-tailored feedback was not significantly different from the values-tailored feedback (p = .343). See Appendix C for the full regression tables and estimated marginal means tables[1].

Study 2. Since the information-tailored feedback condition utilized in [25] was consistently the most effective, we conducted a follow-up study to compare this condition to a more neutral reference group (instead of the adversarial statement condition). For this subsequent study, we tested the impact of both the information-tailored and adversarial statement conditions relative to a neutral control condition, where participants had a non-election-related conversation with GPT about their preference for cats or dogs.

The information-tailored feedback type again significantly decreased 2020 Denialism by 3.56 scale points, or 5%, relative to the more neutral, non-election-related control condition where participants had a conversation about their preference for cats or dogs (b = -3.56, β = -.11, p = .013, d = -.20, Δ% = -5.34%). The adversarial statement condition showed a non-significant negative effect (p = .113). The information-tailored feedback was not statistically different from the adversarial statement treatment (p = .706).

Importantly, relative to the neutral control, Claim Certainty again significantly decreased with information-tailored feedback (b = -4.43, β = -.24, p = .001, d = -.23, Δ% = -4.11%), while the adversarial statement showed no significant effects (p = .870).

Belief change is mediated by reduced certainty in articulated claims

Additionally, we conducted post-hoc mediation models to understand the intervention's impact on both 2020 and 2024 denialism, and the extent to which it was mediated by a change in the specific claim they made. The models were estimated using maximum likelihood, bootstrapped standard errors (10,000 bootstrap samples), and the Bollen-Stine test for model fit. Information-tailored feedback, relative to both reference categories (adversarial statement and the neutral control), showed full mediation effects across both the 2020 and 2024 denialism models in both studies, with changes in claim certainty significantly mediating the effects of the detailed intervention on both 2020 and 2024 denialism (p's < .005; see Appendix E).

[1] There were also no significant differences in intervention effectiveness based on interface type when we looked at our preregistered interaction effects (p's > .087). However, trendwise, the conversations within the chatbot interface resulted in marginally higher effectiveness (see Supplementary Table C1 and C2 for preregistered models and the models controlling for interface type (reported above), respectively). As a result, the chatbot interface was the only one used for Study 2.

Robustness and analytic checks

To assess the sensitivity of results to analytic choices and missing outcomes, we conducted three additional checks. First, because the primary analyses focus on participants whose open-ended responses expressed a specific election-fraud conspiracy (coded via NLP, consistent with prior work, such as [25]), we re-estimated models without the NLP-based restriction and observed similar effect sizes in the full eligible sample (N = 3,006), indicating that results were not impacted by NLP screening.

Second, to account for missing post-experiment scores and some evidence of differential attrition in study 1, we also employed a conservative imputation method by substituting pre-experiment scores for the missing post-experiment scores to explore intention-to-treat outcomes. The results were consistent with the primary models when including these imputed values, suggesting that the missing data did not systematically bias the observed effect

Third, we fitted a Bayesian linear regression with weakly informative priors centered on no effect. Posterior estimates again favored belief reductions of around 4 to 5 pp for the information-tailored (and values-tailored, in Study 1) conditions, while the adversarial statement condition showed smaller and less certain effects. Full model outputs are reported in the Supplementary Materials.

Supplementary Appendices B–D report effects on broader election attitudes (general denialism, expected legitimacy of the 2024 election, and affective polarization) and voting intentions for the 2024 election. Across these outcomes, we find limited evidence of reliable intervention effects. Consistent with this, interventions did not significantly change self-reported likelihood of voting in 2024, which was high overall (Likelihood of voting in 2024 = 92.65%, Supplementary Table C6).

Heterogeneity: Who changes their beliefs?

We also conducted post-hoc, exploratory analyses to investigate differential treatment effects, particularly among participants with deeply rooted beliefs, by fitting Generalized Additive Models (GAMs). The smoothed interaction showed a primarily linear relationship, and all interactions were non-significant (see Fig. 2A & Supplementary Table D1-D2). As such, we report linear regressions instead of GAMs for all analyses.

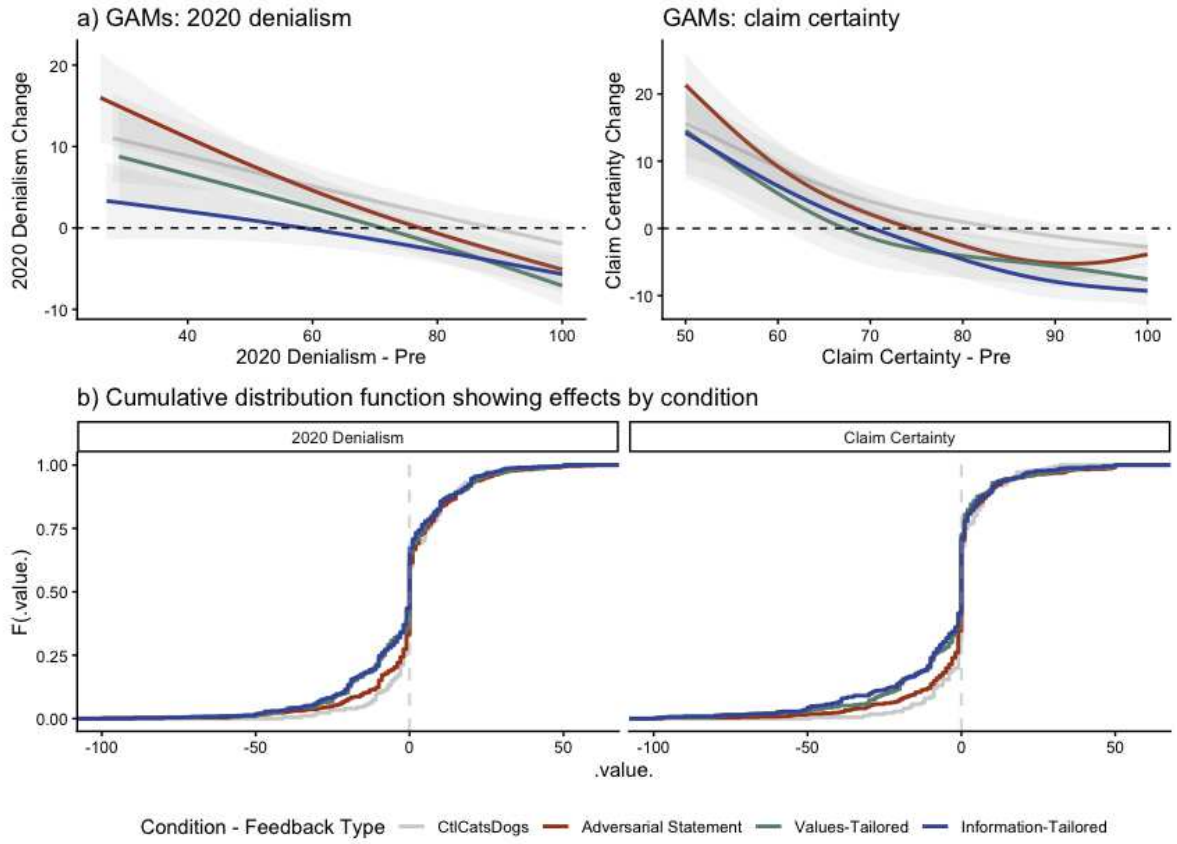
Die-Hards vs. Non-Die-Hards. GAMs highlighted the trend-wise non-linearity of effects, where those with stronger baseline denialism scores had stronger reductions in denial beliefs. Across both studies, over a third of participants (33.40%) stated that they were 100% certain about the 2020 election conspiracy they had stated (referred to as "die-hards" for subsequent analysis). We dichotomized participants into "die-hards" versus "non-die-hards", as recommended by Green[31], to understand differential effects by baseline levels.

For this subset of the population, when examining treatment effects, there were slightly higher, significant effects across election denial outcomes for those with stronger beliefs (p's < .084). For "die-hards", we saw statistically significant changes ranging from 6 to 7% across both studies in 2020 denialism. Meanwhile, approximately 15% of this population had no change in 2020 denial beliefs, and 23% had no change in the specific claim they made (claim certainty) (see Supplementary Table D3 for interaction effects, and Supplementary Table D4 for die-hard vs. non-die-hard effects).

Cumulative Effects. To assess the distribution of effects, we utilized a cumulative distribution function (CDF). Over a third of the participants (38% on average) had post-intervention decreases in claim certainty across

all conditions. While 31% of participants in the denialism condition had reductions in denial beliefs, 35% decreased in the adversarial statement condition, 41% in the values-tailored condition, and 42% in the information-tailored condition (see Fig. 2B). We see similar trends for 2020 Denialism.

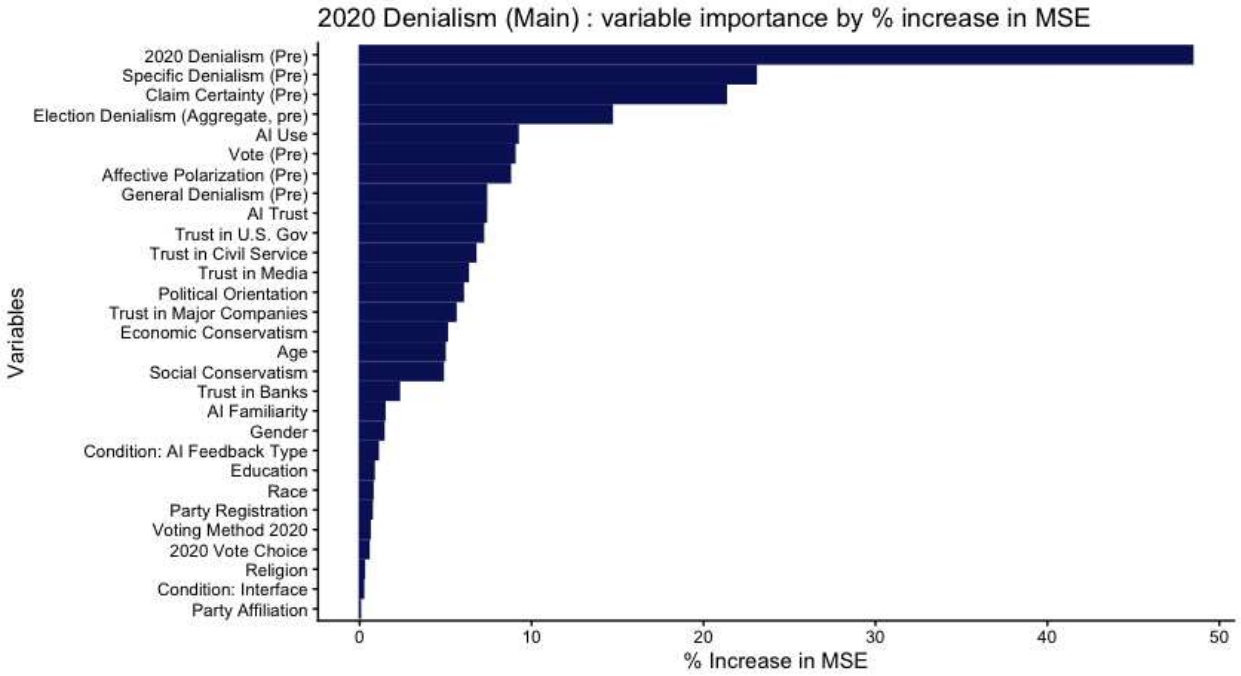
Fig. 2: Change in beliefs as a function of baseline conviction, by condition.



Notably, 74% of participants showed a decrease in at least one election denial outcome, with up to 76.94% of participants' beliefs improving in the information-tailored condition and 75.95% in the values-tailored condition.

We also aimed to understand which covariates most significantly drove effects. Through random forest models, we found that the key outcomes driving changes in election denial beliefs were prior denial beliefs, AI use and trust, voting likelihood, and institutional trust (see Fig. 3).

Fig. 3: Random forest variable importance for predicting post-intervention 2020 Denialism.

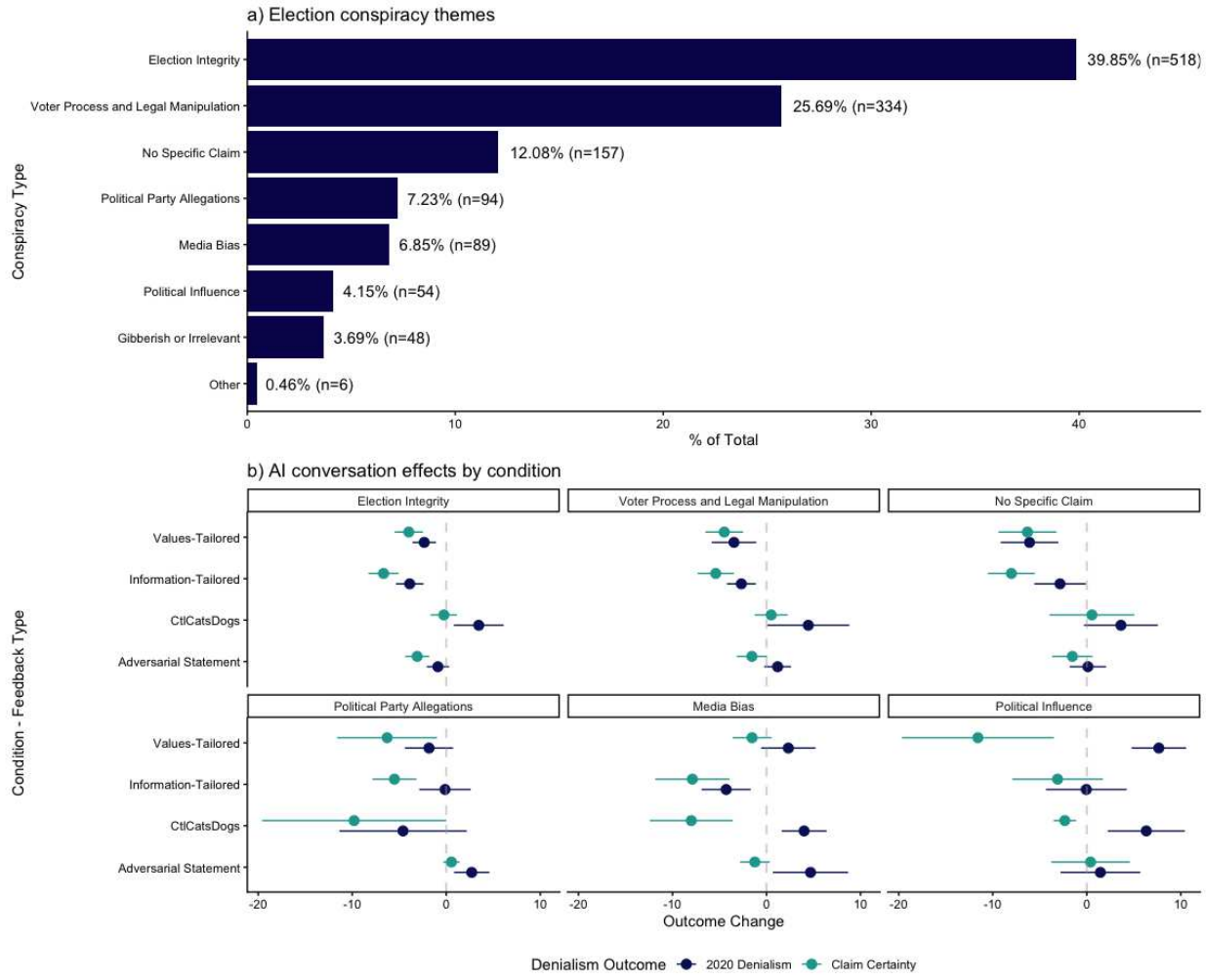


Descriptive content analyses

Additionally, using NLP, we explored the underlying themes in election concerns for those who stated a conspiracy. Thematically, most participants' election conspiracies focused on election integrity (40%), or concerns around the election process being manipulated, rigged, or fraudulent, and suggesting widespread fraud or manipulation of the election. This was followed by claims about voter eligibility, voter fraud, or procedural manipulation (26%), then more general expressions of election fraud without specific reasons or details (12%). The information-tailored treatment retained the strongest efficacy across most themes (Fig. 4).

Otherwise, sentiment analysis was conducted using R's sentimentr package to understand how sentiment changed throughout conversation rounds, and if they were primarily negative or positive. For the first round, where they made an initial statement on the conspiracy they believed, participant sentiment was more negative. However, they moved toward neutrality throughout the conversation (see Supplementary Fig. C3).

Fig. 4: Election denialism themes and condition effects by theme.



Discussion

Through two preregistered studies with Republicans who endorsed 2020 election-fraud conspiracies, we found that, in accordance with our hypotheses, corrective conversations with an AI large language model significantly reduce belief in election denialism claims (i.e., Claim Certainty, 2020 Denialism). Existing research shows that election-fraud beliefs are difficult to change. Within this context, the fact-checking interventions reducing denialism by 5-8% is notable. While effects are relatively modest, they meaningfully move beliefs in a population known to hold durable conspiratorial views. Seeing these effects in spite of this shows that interventions like this may provide a unique avenue for changing these harder-to-change, inaccurate beliefs, like those around election fraud.

Although there is evidence that motivational message matching can increase persuasion in other domains [32], we did not observe added benefits to tailoring arguments based on the participant's values, which is in line with previous work on the limited effects of micro-targeting [30]. Overall, it showed weaker and sometimes non-significant effects across election denialism outcomes. This potentially suggests that AI-based, specific, relevant facts may matter more than hyper-personalized corrections framed around a participant's identity within this

context (as in [26]). Given the limited additional value to further personalization, there is an opportunity to conduct further subgroup analysis to better understand which tailoring attributes, or types of values, to effectively tailor on, and for which subset of election fraud conspiracy believers (e.g., by using motivational matching methodology).

We also found some evidence that the less fact-centric adversarial statement condition, which was designed to avoid counterargument, produced small decreases in some election denialism outcomes in Study 1, but not relative to a neutral control in Study 2. This may be due to slight modifications being made to the GPT-fed prompt in study 2 (see Methods and Supplementary Table B2), or slight population differences. This condition may potentially have deflated effect sizes for study 1 by also acting as a treatment intervention and debunking election conspiracies. It was unclear whether the adversarial statement condition was somewhat effective due to successful persuasion, demand characteristics, or other considerations. However, this provides an initial indication supporting dual process models' propositions that updating beliefs may not necessarily need extensive fact-checking, but a central route-based nudge to not focus on inaccurate beliefs. Regardless, our results ultimately point toward claim-level, tailored fact-checking most effectively driving changes in fraud beliefs.

Furthermore, we see some evidence that it was significantly more effective for those with more deeply entrenched beliefs, or "die-hards." Bayesian belief updating theories suggest that individuals update beliefs as a function of their priors. Within this context, it is expected that beliefs are hardest to change for those with more deeply embedded priors. However, we see stronger effects more consistently for this subset of the population. This effect is notable because it shows that these types of belief-tailored, AI-based interventions can be extremely powerful for the most vulnerable subsets of the population, contrary to traditional Bayesian theory.

Furthermore, these AI conversations show stronger effects for the specific beliefs participants had (i.e., claim certainty), versus more general election legitimacy skepticism (i.e., 2020/2024/general denialism). This could be because we target more deeply embedded or core beliefs. Otherwise, mediation analyses suggest that the interventions primarily worked by lowering certainty in participants' own articulated claims, which in turn predicted lower 2020 denialism. This is consistent with argument-specific updating, rather than broad shifts in generalized trust or global attitudes.

This mechanism is theoretically important. It suggests that tailoring arguments to an individual's specific beliefs is most effective when grounded in the participant's actual fraud claim, enabling the model to address inaccuracies directly. It also provides some clarification on why values-tailoring did not outperform information-tailoring. The psychological appeal may come from targeting the belief, not from matching the message to personal values.

Limitations. There were several limitations. Firstly, we have an unclear understanding of why values-tailored messaging was not at least as effective as the information-tailored treatment. Additional research should explore the contexts in which tailoring can be effective, and which types. AI as a persuasion agent may be less effective for some topics. It may necessitate more research on belief formation or training on how to collect information. Additionally, though people received responses from AI about their election denial beliefs, issues such as selective exposure could have resulted in participants choosing not to engage with AI in a meaningful way or to leave the experiment, directly impacting efficacy. This is a constraint regardless of the context in which this type of intervention is deployed, but it should still be considered.

Otherwise, there have been potential concerns, with easier AI access, around the potentially paternalistic nature of some of these interventions (as flagged by Zettler & Strandsbjerg[33]), and success being influenced by the innate biases of LLMs, and the programming/prompt injection of the researchers. Likewise, there are limitations to replicability due to the black box that is LLMs, the ever-evolving nature of the code, and the underlying programming that determines how they respond

Finally, this work was also done leading up to the 2024 US Presidential election and only focuses on Republicans. With the change in political party leadership to a Republican president, there's some existing research showing that conspiratorial beliefs are linked to one's preferred party in power. This may be less applicable and effective within the current context, due to lower baseline skepticism for Republicans since their aligned party is now in power.

Future Directions. Future research should aim to further explore argument tailoring or to understand which tailoring facets are most effective within the context of election fraud beliefs. Additionally, research should further explore the mechanistic differences and how tailored, AI-based debunking can be more effective for populations. Otherwise, there is value in understanding how AI feedback compares to static feedback (e.g., everyone gets the same feedback) within this and other contexts. Otherwise, there may be value in further refining GPT prompts and argumentation strategies (arguing to learn instead of teach [34], empathy [35,36], more perspective-taking [37], learnings from conflict resolution and negotiation strategies [38], shorter responses, and other strategies that show evidence of efficacy across the literature). Furthermore, there is an opportunity to explore and test these and similar interventions in the field (e.g., platforms more prone to conspiracy theories).

Conclusion. Our studies show evidence that AI-driven conversations can reduce specific inaccurate election-fraud beliefs, with the strongest effects emerging when models provide detailed, claim-targeted evidence. While we see significant reductions, values-based tailoring did not result in added benefit in this domain, suggesting that fact-checking, rather than identity-matching, may be the most reliable pathway for belief correction in politically charged contexts, and that additional research on how to hyper-personalize corrections may be valuable. As interactive AI becomes increasingly integrated into public information environments, understanding when and how it can responsibly reduce misinformation will be essential for safeguarding democratic processes such as voting and perceptions of legitimacy.

Social platforms like X (Twitter) are rapidly integrating generative AI (e.g., Grok) and other tools like Community Notes to offer real-time context and corrections. These findings provide insight into the potential utility of AI-based tools and offer insight into the role they could play in social media environments by providing credibility-enhancing or counter-attitudinal information dynamically.

Method

All studies were deemed minimal risk and exempt by the MIT Committee on the Use of Humans as Experimental Subjects (protocol # E-5984), with supplementary approval by Duke University for analysis of existing data (protocol # 2025-0390). Participants for both studies were recruited through the survey-taking platform, Lucid[2], and compensated for participation. We limited recruitment to US adult Republicans, based on public opinion data showing that election denialism is more prevalent among Republicans [39].

[2] Lucid is a platform known for non-compliance and inattention. Despite this, we still see significant intervention effects.

Experimental Design

For **Study 1**, a 3 x 2 factorial design was used, where participants were randomly assigned to one of three AI-led corrective conversation interventions (information-tailored, values-tailored, or adversarial statement) and one of two conversational interfaces (Qualtrics or chatbot). In all conversations, GPT disputed the specific claims of election fraud, and in the information-tailored and values-tailored conditions, fact-checking was used for these refutations. For **Study 2**, participants were assigned to one of three conditions (information-tailored, adversarial statement, or the cats and dogs control). In all conditions, participants engaged in a three-round conversation with GPT-4 Turbo or 4o (adversarial statement only) about the election conspiracy they initially stated.

See Table 1 for an overview of each condition. Figure B3 shows the visual differences in both interfaces. All model prompts and models used during the experiments are provided in Supplementary Table B2.

Table 1: Overview of experimental conditions.

Condition	Variant	Overview	Studies Used
Information-Tailored	Message	Using procedures and slightly modified prompts from Costello et al. (2024), updated to focus on election conspiracies. GPT was prompted to effectively provide corrective information and persuade participants that the elections were not rigged.	1 & 2
Values-Tailored	Message	Using procedures and prompts from Costello et al. (2024), as in the information-tailored feedback condition; however, unlike this treatment, the GPT prompt also included detailed information about the participants' core values. In particular, we used their open-ended responses to a question on their values (i.e., tell us what you care about and want your life to stand for). GPT was prompted to organically tailor arguments to their values (without compromising the efficacy of its persuasion), but not to make it obvious they were aware of the users' values. Due to limited evidence of effectiveness relative to information-tailored feedback, this was only used for Study 1.	1
Adversarial Statement	Message	This condition aimed to evaluate the extent to which factual information drives the effect. Thus, we instructed GPT-4 to persuade without making any factual claims or using rationality, counterevidence, critical thinking, alternative explanations, and/or reason-based persuasion strategies. Examination of model responses using NLP showed that in 84.71% of the conversations, this was upheld, and evidence was not used. On average, 0.91 pieces of evidence were provided per conversation within this condition (0.87 in study 1 and 1.07 in study 2). Thus, in some instances, facts and evidence were delivered – albeit to a lesser extent than in the treatment conditions (with 2.84 to 3.20 pieces of evidence per conversation).	1 & 2
Cats & Dogs Conversation	Message	A conversation with GPT-4o where the AI is instructed to engage in a debate about whether cats or dogs are better companions.	2
Qualtrics Interface	Interface	Conversations were conducted through Qualtrics' interface as question text, as was done in Costello et al. (2024).	1
Chatbot Interface	Interface	Conversations were conducted in a chatbot-like interface using prompts from Costello et al. (2024). The chatbot was embedded into Qualtrics' interface using JavaScript. See vegapunkdoc.dev for additional details on the chatbot interface.	1 & 2

Procedure

Survey Flow. At the beginning of the survey used for both studies, participants were asked to provide consent, complete a Captcha check, and complete an initial question sharing their primary values (e.g., "In a few sentences, what's most important to you in life? What values guide how you live and interact with others?"). They were then asked demographic questions (e.g., age, race, gender, religiosity), about political practices and perceptions (e.g., political affiliation, voting behaviors and intentions, institutional trust, and affective polarization), and about knowledge of and trust in AI, and frequency of use. The full survey can be viewed at https://osf.io/2cfkv/.

Participants were asked to state a specific election conspiracy they believe in, which was fed to a version of GPT-4 within Qualtrics, then summarized by GPT. Participants were then asked to rate their belief in the summarized conspiracy's veracity using a scale that ranged from 0 [Definitely false] - 100 [Definitely true] (measuring **Claim Certainty**, a primary dependent variable). Participants were also asked a series of election denialism questions. A pre-post design was used, where all participants were asked the same primary outcome questions again after completing the intervention (see Supplementary Table B1 and the *Outcomes* subsection for details on the main measures).

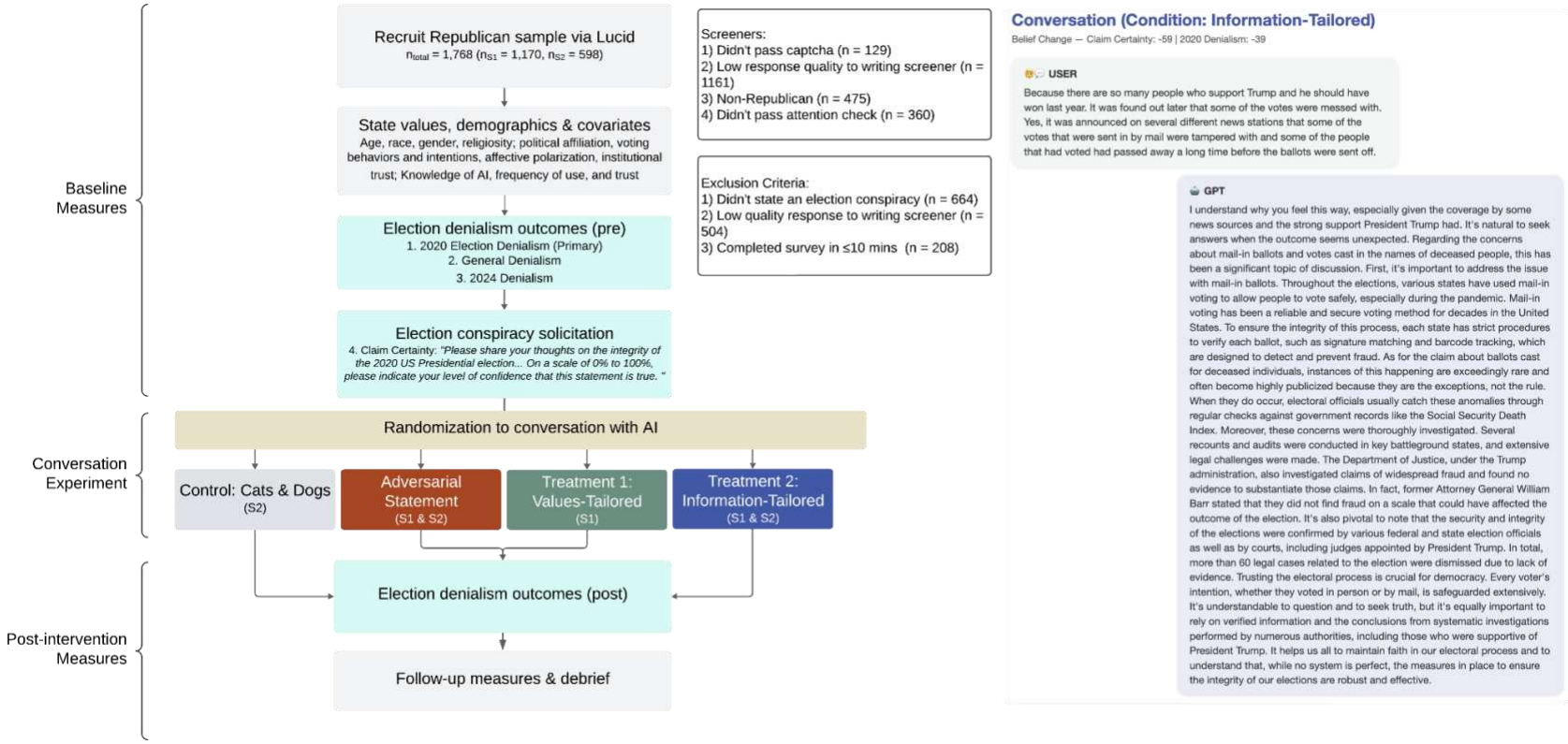
Following these pre-treatment denialism measures, participants were informed they would be conversing with an advanced AI (see Supplementary Table B2 for the researchers' backend GPT prompts for each condition). They then engaged in the GPT conversation intervention. At the end of the survey, all participants were debriefed and informed about the limitations and constraints of generative AI models.

AI Conversation Intervention. Within all experimental conditions, participants engaged in a multi-round conversation with GPT. To facilitate this real-time interaction within the Qualtrics survey platform, we used JavaScript to call OpenAI's Chat Completions API. We dynamically provided participant-specific information about their election beliefs into the model's instructions about the claims they stated and their initial belief levels. See Table S5 for the GPT model and prompts used for each experimental condition.

Conversations were initiated by GPT's initial response to the claim they made or by asking about their preference for cats or dogs (the control condition). This was followed by alternating user and AI messages for 3 rounds. There was no token limit placed on the AI's responses. On average, GPT responses were 206 words long, while participant responses were 44.3 words long, with conversations taking place over 8.42 minutes (Cats & Dogs: 4.49 mins; 71.0 words; Adversarial Statement: 5.85 mins, 62.4 words; Values-Tailored: 10.9 mins, 165 words; Information-Tailored: 9.76 mins, 168 words).

See Fig. 5 for an overview of the study flow, participants screened out and excluded from analysis, and Supplementary Table A3 for an overview of attrition by experimental condition and study.

Fig. 5: Study flow and experimental design.



Outcomes. The primary preregistered election denialism outcome was 2020 Denialism, or participants' beliefs that fraud changed the outcome of the 2020 U.S. presidential election (i.e., "How likely do you think it is that fraud (e.g., voter fraud, rigging, tally manipulation) changed the outcome of the 2020 US Presidential election?") on a 0-100 scale. We also pre-registered that we would create a 7-item general election denialism scale if items were correlated with $r > .80$; however, all inter-item correlations were $r < .80$. Thus, for simplicity, we do not include analyses of these variables in the primary results. Details on these measures can be found in Appendix B.

As a secondary outcome, we pre-registered Claim Certainty, which was focused on person-specific election conspiracy beliefs. Participants were asked the following open-ended question: "Please share your thoughts on the integrity of the 2020 US Presidential election." Following each open-ended response, GPT-4 summarized their question as a declarative, high-level statement of belief (e.g., "The 2020 US Presidential election had some irregularities, supported by investigations and media reports, which make it hard to believe in the legitimacy of the vote counts that led to Biden's late surge and victory, despite Trump's substantial leads."). Participants were shown this summary and asked, "On a scale of 0% to 100%, please indicate your level of confidence that this statement is true".

Participants

Across both studies, the total sample comprised 1,768 participants. Study 1's final sample consisted of 1,170 participants, with 174 to 214 participants assigned to each experimental condition (information-tailored, values-tailored, or adversarial statement, and chatbot or Qualtrics)[3]. The sample comprised 694 (59.32%) females and 456 (38.97%) males, with participants ranging in age from 18 to 98 years old (Mean = 56.29, SD = 17.14). Participants were primarily White (92.03%), Black (3.28%), and Asian (2.13%). All participants identified as Republican (Mean = 5.21, SD = .76) on a 6-point scale (1 = Strongly Democratic, 6 = Strongly Republican). There were no significant differences in demographics and political behaviors by condition.

Participants who believed an election conspiracy exhibited several differences in demographics and beliefs compared to the full sample of 4,320 participants that entered the survey (i.e., older, less educated, female, more religious, stronger affective polarization, more conservative, Trump supporters, used mail-in voting for the 2020 elections, to vote in 2024 elections; p's < .001). Supplementary Table A1 shows demographic differences between the study 1 and study 2 samples, and Supplementary Table A2 shows the differences between the full and final samples.

Attrition rates (the percentage of people who discontinued the survey during the intervention, or those with a pre-intervention claim certainty score but no post-intervention score) ranged from 3.55% (n = 7) in the Adversarial Statement + Qualtrics interface condition to 15.05% (n = 31) in the Information-Tailored + Chatbot interface condition. Chi-squared tests were used to check if failure to complete the survey varied based on treatment assignment. There were significant differences in attrition based on condition assignment ($\chi^2(5) = 25.72$, $p < .001$), with both interface type ($\chi^2(1) = 13.26$, $p < .001$) and feedback type ($\chi^2(2) = 8.36$, $p = 0.015$) significantly impacting attrition (see Supplementary Table A3 for attrition rates by condition). Given this differential attrition, we include a post-hoc robustness check using baseline-carried-forward imputation in which

[3] We aimed for balanced condition assignment across conditions, but due to differential attrition and exclusion criteria, final sample sizes varied across conditions.

we assume that all participants who drop out had no reduction in belief (see Supplementary Table C5 and the results section).

For Study 2, the final sample consisted of 598 participants. The sample comprised 329 (55.20%) females and 257 (43.12%) males, and participants ranging from 18 to 93 years old (Mean = 55.16, SD = 18.42). Participants were primarily White (90.83%), Black (5.54%), and Asian (1.38%). All participants identified as Republican (Mean = 5.18, SD = .73) on a 6-point scale. There were no significant differences in demographics or political behaviors across conditions. The full demographic details and political behaviors are shown in Appendix A.

Study 2 attrition rates ranged from 12.38% (n = 25) in the adversarial statement condition to 14.05% (n = 34) in the information-tailored condition and 14.89% (n = 35) in the cats & dogs control condition. Chi-squared tests were used to check if failure to complete the survey varied based on treatment assignment. We did not see significant differences in attrition by condition ($\chi^2(2) = 0.59$, $p = .744$).

Data & Code Availability

This study's design, research questions, and analyses were pre-registered. Our pre-registration, data, surveys, and code can be found on OSF (Study 1: https://osf.io/rq3fm, preregistered on Jul 29, 2024; Study 2: https://osf.io/pgdzw, preregistered on Sep 25, 2024).

References

1. Lee, T. The global rise of "fake news" and the threat to democratic elections in the USA. *Public Adm. Policy* **22**, 15–24 (2019).

2. Mauk, M. & Grömping, M. Online Disinformation Predicts Inaccurate Beliefs About Election Fairness Among Both Winners and Losers. *Comp. Polit. Stud.* **57**, 965–998 (2024).

3. Bell, P. Public Trust in Government: 1958-2025. *Pew Research Center* https://www.pewresearch.org/politics/2025/12/04/public-trust-in-government-1958-2025/ (2025).

4. Hartmann, D., Wang, S. M., Pohlmann, L. & Berendt, B. A systematic review of echo chamber research: comparative analysis of conceptualizations, operationalizations, and varying outcomes. *J. Comput. Soc. Sci.* **8**, 52 (2025).

5. Rhodes, S. C. Filter Bubbles, Echo Chambers, and Fake News: How Social Media Conditions Individuals to Be Less Critical of Political Misinformation. *Polit. Commun.* https://www.tandfonline.com/doi/full/10.1080/10584609.2021.1910887 (2022).

6. Elsner, M., Atkinson, G. & Zahidi, S. *Global Risks Report 2025*. (Forum Publishing, Cologny/Geneva, Switzerland, 2025).

7. Thomas, E. F. *et al.* Conspiracy beliefs and democratic backsliding: Longitudinal effects of election conspiracy beliefs on criticism of democracy and support for authoritarianism during political contests. *Polit. Psychol.* **n/a**, (2025).

8. Saad, L. Partisan Split on Election Integrity Gets Even Wider. *Gallup.com* https://news.gallup.com/poll/651185/partisan-split-election-integrity-gets-even-wider.aspx (2024).

9. Norris, P. & Inglehart, R. *Cultural Backlash: Trump, Brexit, and Authoritarian Populism*. (Cambridge University Press, 2019).

10. Sances, M. W. & Stewart, C. Partisanship and confidence in the vote count: Evidence from U.S. national elections since 2000. *Elect. Stud.* **40**, 176–188 (2015).

11. Kulke, S. 38% of Americans lack confidence in election fairness. https://news.northwestern.edu/stories/2020/12/38-of-americans-lack-confidence-in-election-fairness (2020).

12. Masaru, N. Presidency of Donald Trump and American Democracy: Populist Messages, Political Sectarianism, and Negative Partisanship. *Asia-Pac. Rev.* **28**, 80–97 (2021).

13. Eggers, A. C., Garro, H. & Grimmer, J. No evidence for systematic voter fraud: A guide to statistical claims about the 2020 election. *Proc. Natl. Acad. Sci.* **118**, e2103619118 (2021).

14. Alvarez, R. M., Hall, T. E. & Hyde, S. D. *Election Fraud: Detecting and Deterring Electoral Manipulation*. (Bloomsbury Publishing USA, 2009).

15. Cottrell, D., Herron, M. C. & Smith, D. A. Vote-by-mail Ballot Rejection and Experience with Mail-in Voting. *Am. Polit. Res.* **49**, 577–590 (2021).

16. Mosleh, M., Martel, C., Eckles, D. & Rand, D. Perverse Downstream Consequences of Debunking: Being Corrected by Another User for Posting False Political News Increases Subsequent Sharing of Low Quality, Partisan, and Toxic Content in a Twitter Field Experiment. in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* 1–13 (ACM, Yokohama Japan, 2021). doi:10.1145/3411764.3445642.

17. Walter, N. & Murphy, S. T. How to unring the bell: A meta-analytic approach to correction of misinformation. *Commun. Monogr.* **85**, 423–441 (2018).

18. Bailard, C. S., Porter, E. & Gross, K. Fact-checking Trump's election lies can improve confidence in U.S. elections: Experimental evidence. *Harv. Kennedy Sch. Misinformation Rev.* https://doi.org/10.37016/mr-2020-109 (2022) doi:10.37016/mr-2020-109.

19. Carey, J. M. *et al.* The Narrow Reach of Targeted Corrections: No Impact on Broader Beliefs About Election Integrity. *Polit. Behav.* **47**, 737–750 (2025).

20. Bolsen, T. & Palm, R. Motivated Reasoning and Political Decision Making. in *Oxford Research Encyclopedia of Politics* (2019). doi:10.1093/acrefore/9780190228637.013.923.

21. Claassen, R. L. & Ensley, M. J. Motivated Reasoning and Yard-Sign-Stealing Partisans: Mine is a Likable Rogue, Yours is a Degenerate Criminal. *Polit. Behav.* **38**, 317–335 (2016).

22. Doell, K. C., Pärnamets, P., Harris, E. A., Hackel, L. M. & Van Bavel, J. J. Understanding the effects of partisan identity on climate change. *Curr. Opin. Behav. Sci.* **42**, 54–59 (2021).

23. Kahan, D. M., Jenkins-Smith, H. & Braman, D. Cultural cognition of scientific consensus. *J. Risk Res.* **14**, 147–174 (2011).

24. Brashier, N. M., Pennycook, G., Berinsky, A. J. & Rand, D. G. Timing matters when correcting fake news. *Proc. Natl. Acad. Sci.* **118**, e2020043118 (2021).

25. Costello, T., Pennycook, G. & Rand, D. Durably reducing conspiracy beliefs through dialogues with AI. https://www.science.org/doi/abs/10.1126/science.adq1814 (2024).

26. Costello, T. H., Pennycook, G. & Rand, D. Just the facts: How dialogues with AI reduce conspiracy beliefs. Preprint at https://doi.org/10.31234/osf.io/h7n8u_v1 (2025).

27. Lin, H. *et al.* Persuading voters using human–artificial intelligence dialogues. *Nature* **648**, 394–401 (2025).

28. Argyle, L. P. *et al.* Testing theories of political persuasion using AI. *Proc. Natl. Acad. Sci.* **122**, e2412815122 (2025).

29. Hackenburg, K. *et al.* The levers of political persuasion with conversational artificial intelligence. *Science* **390**, eaea3884 (2025).

30. Hackenburg, K. & Margetts, H. Evaluating the persuasive influence of political microtargeting with large language models. *Proc. Natl. Acad. Sci.* **121**, e2403116121 (2024).

31. Green, J. A. Too many zeros and/or highly skewed? A tutorial on modelling health behaviour as count data with Poisson and negative binomial regression. *Health Psychol. Behav. Med.* **9**, 436–455 (2021).

32. Joyal-Desmarais, K. *et al.* Appealing to motivation to change attitudes, intentions, and behavior: A systematic review and meta-analysis of 702 experimental tests of the effects of motivational message matching on persuasion. *Psychol. Bull.* **148**, 465–517 (2022).

33. Zettler, I. & Strandsbjerg, C. F. Personalized interventions. *Curr. Opin. Psychol.* **66**, 102147 (2025).

34. Fisher, M. & Keil, F. C. The Trajectory of Argumentation and its Multifaceted Functions.

35. Lee, S., Kim, S. & Bae, J. The Effects of empathy and personalization on chatbot usage intention: A structural equation modeling analysis based on the elaboration likelihood model (ELM). *Int. J. Adv. Smart Converg.* **14**, 78–85 (2025).

36. Hangartner, D. *et al.* Empathy-based counterspeech can reduce racist hate speech in a social media field experiment. *Proc. Natl. Acad. Sci.* **118**, e2116310118 (2021).

37. Voelkel, J. G. *et al.* Megastudy testing 25 treatments to reduce antidemocratic attitudes and partisan animosity. *Science* **386**, eadh4764 (2024).

38. Movius, H. The Effectiveness of Negotiation Training. *Negot. J.* **24**, 509–531 (2008).

39. Stewart III, C. Public Opinion Roots of Election Denialism. SSRN Scholarly Paper at https://doi.org/10.2139/ssrn.4318153 (2023).

Figure Legends

Fig. 1: Denialism changes by condition & study.

Points display mean estimated condition effects with standard errors for changes in (left) 2020 Election Denialism and (right) Claim Certainty. Effects are estimated separately by study, comparing each condition to the study-specific reference group (Study 1: Adversarial Statement; Study 2: Cats & Dogs control), controlling for mean-centered baseline beliefs (and interface condition in Study 1). Negative values indicate reductions in denialist beliefs. Colors denote Study 1 (dark blue) and Study 2 (teal). Error bars indicate standard errors.

Fig. 2: Change in beliefs as a function of baseline conviction, by condition.

Top panel: Change in 2020 Denialism (post–pre) plotted against baseline 2020 Denialism (left) and Change in Claim Certainty (post–pre) plotted against baseline Claim Certainty (right). Lines show linear fits with 95% CIs by condition. Negative values on the y-axis indicate reductions in denialism/certainty. Across panels, higher baseline conviction is associated with larger decreases following Information-Tailored feedback; Values-Tailored shows similar but smaller slopes; Adversarial Statement is weaker and less consistent.

Bottom panel: Cumulative distribution functions of (left) change in 2020 Denialism and (right) change in Claim Certainty (post–pre). Curves further left reflect larger decreases in fraud beliefs. Information-tailored and values-tailored conditions show left-shifted distributions relative to controls, indicating a higher share of participants with meaningful reductions; Adversarial Statement shows smaller shifts.

Fig. 3: Random forest variable importance for predicting post-intervention 2020 Denialism.

Random forest model estimating the relative contribution of pre-intervention characteristics, political attitudes, institutional trust, demographic variables, and experimental condition to post-intervention 2020 Denialism. Bars represent the percentage increase in mean squared error (%IncMSE) when each predictor is permuted, with larger values indicating greater predictive importance. Pre-intervention election fraud belief strength, political identity measures (e.g., partisan affiliation, affective polarization), and institutional trust variables emerged as the strongest predictors of post-intervention certainty. Experimental conditions contributed meaningfully but less strongly than participants' prior beliefs. Only predictors with positive values (%IncMSE > 0) are displayed.

Fig. 4: Election denialism themes and condition effects by theme.

Top panel: Distribution of participant-stated conspiracy themes derived via NLP classification of open-ended claims (e.g., Election integrity, Voter process/legal manipulation, Media bias). Bars show the percent of all coded claims. Bottom panel: Condition effects (point estimates with standard errors) on change in 2020 Denialism and Claim Certainty within each theme. Negative values indicate reductions. Information-tailored feedback generally produced the most consistent decreases in election fraud beliefs across major themes.

Fig. 5: Study flow and experimental design.

Overview of the study flow, reflecting capturing baseline measures, randomization, the AI conversation interventions, and post-intervention outcomes. Republicans who endorsed an election-fraud claim on Lucid were recruited (total N = 1,768; Study 1, n = 1,170; Study 2, n = 598). The second panel shows a sample conversation between the participant and GPT where the participant was assigned to the adversarial statement condition.

Table 1: Overview of experimental conditions.

Note. Overview of all experimental variations used in Studies 1 and 2. Message-based conditions manipulated the content and style of GPT's responses: Information-tailored feedback provided detailed, claim-specific factual rebuttals; values-tailored feedback additionally integrated each participant's stated core values (Study 1 only); the adversarial statement attempted to persuade without relying on facts, evidence, or reason-based strategies; and the cats & dogs condition served as a neutral, non-election-related control conversation. Interface-based conditions manipulated the presentation format of the AI dialogue: a text-based Qualtrics Interface (Study 1 only) versus an embedded Chatbot Interface (Studies 1 and 2)

Author Contributions

SH, TC, GP, & DR conceptualized the study and designed the research. SH & TC performed research. SH analyzed data. SH wrote the paper. SH, TC, GP, & DR reviewed and edited the final manuscript.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- AIElectionsPaperSupplement.pdf