

When Information Affects Attitudes: The Effectiveness of Targeting Attitude-Relevant Beliefs

Yamil Velez, Patrick Liu*, and Scott Clifford†

June 6, 2025

Abstract

Do citizens update strongly held beliefs when presented with incongruent information? Does belief change then precipitate change in downstream attitudes? A persistent puzzle in the study of public opinion is why corrective information frequently produces negligible attitude change despite reliably influencing beliefs. We argue that this duality reflects belief relevance—the extent to which specific beliefs bear on attitudes. Using large language models (LLMs), we elicit deeply held issue attitudes and “focal beliefs” described as central to those attitudes. We then randomly assign participants to receive either an LLM-generated counterargument targeting (1) their focal belief, (2) an attitude-relevant but unmentioned belief (“distal belief”), or (3) a placebo. In experiments with two large online convenience samples ($N_1 = 2,468$; $N_2 = 1,910$), we show that counterarguments successfully decrease both focal and distal belief strength, with effects persisting after one week. More importantly, focal belief counterarguments produce larger and more durable attitude change than distal counterarguments. These findings suggest that corrective information can successfully shift political attitudes when targeting “focal beliefs,” providing new evidence for the role of belief relevance in persuasion.

*Columbia University Department of Political Science

†Texas A&M University Department of Political Science

Introduction

The duties of democratic citizenship place a premium not only on voter learning but on the assimilation of accurate information into reasoned political judgments. Under traditional theories of democratic accountability, voters must be able to convert information about performance into judgments of incumbent governments and cast their votes accordingly (Fearon, 1999). Likewise, the cause for widespread concern about misinformation is not merely that citizens confidently hold false beliefs, but that policy and candidate preferences are formed on their basis (Kuklinski et al., 2000; Hochschild and Einstein, 2015). Education and public information campaigns may supply citizens with accurate beliefs, but their efficacy hinges on citizens translating belief change into attitude change.

Recent studies document that, across varied issues and contexts, interventions that improve citizens' knowledge fail to consistently or substantially change supposedly downstream attitudes—in spite of mounting evidence that people generally update their beliefs in the direction of evidence (Coppock, 2023; Arias et al., 2022; Gerber and Green, 1999). Improving knowledge of government spending and inequality produces inconsistent effects on economic policy preferences (Barnes et al., 2018; Cook, Jacobs and Kim, 2010; Kuziemko et al., 2015). Factual corrections eliminate false beliefs about vaccines without affecting intention to vaccinate (Porter, Velez and Wood, 2022; Helfers and Ebersbach, 2023) and counteract misinformation propagated by political candidates while exhibiting minimal influence on candidate support (Nyhan et al., 2020; Coppock et al., 2023). Efforts to reduce hostility toward immigrants and out-partisans by correcting misperceptions also recover mixed results (Hopkins, Sides and Citrin, 2019; Grigorieff, Roth and Ubfal, 2020; Dias et al., 2024).

A review of the literature reveals little consensus for what this empirical pattern implies about mass political behavior. These findings at least seem to reaffirm the intuition that political attitudes are the sum of citizens' prior knowledge and evaluations and are therefore less elastic than their constituent parts (Barnes et al., 2018). Other popular

interpretations are less sanguine about the benefits of a more informed citizenry. Evidence that learning aggregate economic information facilitates retrospective voting only among politically sophisticated voters suggests some may lack the requisite cognitive skills to make the link between belief and attitude change (Alt, Lassen and Marshall, 2016). Worse still, misperceptions may be the consequence of political attitudes rather than the cause (Hopkins, Sides and Citrin, 2019; Lawrence and Sides, 2014). In this view, beliefs may be little more than post-hoc rationalizations of attitudes conceived in order to shield one’s prior commitments from incongruent evidence, an idea consistent with theories of cognitive dissonance and motivated reasoning (Festinger, 1957; Kunda, 1990; Lodge and Taber, 2013).

Why do interventions so often shift beliefs but fail to move attitudes? Answering these questions, we argue, requires paying more careful attention to the relevance that people themselves assign to beliefs. Factual interventions are ordinarily designed to target beliefs that academics or policy analysts perceive *a priori* as focal to a policy debate or momentarily pertinent due to widely circulating misinformation. Yet to the mass public, a wholly different set of beliefs may matter. For one, the facts citizens deem relevant to a policy debate are shaped by how politicians frame the debate, and thus may differ from what experts prescribe as relevant (Kuklinski et al., 1998). This concern bears resemblance to Converse’s (1964, 8-10) familiar argument that mass ideology is more likely to exhibit “social” than “logical” constraint: voters’ ideas about “what goes with what” tend not to emerge organically as logical wholes, but rather are diffused from elites.¹ Moreover, because relevance depends on politics as well as people’s personal values—which shape the criteria used to determine what information matters (Gerber and Green, 1999, 193)—relevant beliefs may differ with partisan identity and group membership, leaving the false impression that identities bias subsequent attitude change. Even if scholars identify and account for the most salient beliefs within

¹Importantly, the argument and evidence in this paper do not speak directly to Converse’s notion of constraint between idea-elements, which concerns the bundle of attitudes a person holds more so than the set of factual beliefs held about an attitude object. Our findings about belief change facilitating attitude change should not be extrapolated as evidence for dynamic constraint, which describes how change in one attitude might beget change in another.

a population or subgroup, doing so may mask substantial heterogeneity between subjects (Fishbein and Ajzen, 2010, 102), which can mute downstream attitude change. The total effect is that belief change may facilitate little to no attitude change when interventions target beliefs that respondents consider only weakly relevant.

Our argument proceeds in two steps. First, we show that prominent debates about change and rigidity in public opinion address why information fails to change beliefs without explaining how belief and attitude change could diverge. We offer conceptual clarity surrounding the distinction between beliefs and attitudes and summarize how attitudes may be modeled as the weighted sum of beliefs about an attitude object, where the influence of each belief is in direct proportion to its relevance. In doing so, this section illustrates why the concept of relevance enables scholars to reconcile widespread evidence of inconsistent or muted effects of information on attitudes with the standard intuition of belief-attitude models—“a reasonable and consistent link between information about an object and attitude toward that object” (Fishbein and Ajzen, 2010, 99).

To address the longstanding methodological challenge of identifying and targeting a respondent’s personally relevant beliefs, we developed a respondent-tailored design enabled by the large language model (LLM) GPT-4o. In the first stage, respondents engage in dialogue with an AI chatbot about a political attitude they consider deeply important and their reasoning behind that view. GPT then synthesizes a summary attitude, a “focal” belief underlying that attitude, and a second “distal” belief generated without reference to chat input. In the second stage, respondents are randomly exposed to an intervention that targets a focal belief, distal belief, or placebo topic, in each case displaying a paragraph-length counterargument incorporating substantial factual information. This design furthers the development of tailored experimental methods in the political persuasion literature (e.g., Velez and Liu, 2024).

Across the two studies, we provide consistent evidence that belief relevance is crit-

ical to attitude change. Tailored counterarguments were equally effective at shifting focal and distal beliefs. However, the attitudinal effects of focal belief counterarguments were sometimes twice as large, as well as significantly different based on a random effects meta-analysis of twelve attitudinal outcomes. Additionally, in follow-up surveys approximately ten days after initial exposure, we find some persistence of attitudinal effects of the focal belief counterarguments but no significant effects of the distal belief counterarguments. These differences in attitude change despite comparable belief change testify to the importance of accounting for belief relevance at the respondent level. We conclude by exploring additional mechanisms and discussing implications for the studies of persuasion and democratic politics.

Explaining Attitudes' Rigidity Despite Belief Change

The tendency of strongly held attitudes to resist change even in the face of arguments backed by truthful information has long alarmed social scientists (e.g., Festinger, Riecken and Schachter, 1956). Although students of public opinion have written extensively on how citizens respond to new information, recent debates focus on whether and why information fails to alter beliefs, without addressing why attitudes may resist change even as beliefs move.

One widely cited explanation for the rigidity of beliefs and attitudes faults citizens as unwilling to pursue accuracy. Under the theory of motivated reasoning, people are often driven by “directional motives” to reach specific conclusions (Kunda, 1990) and therefore accept and interpret evidence in a selective or skewed fashion (e.g., Bartels, 2002; Taber and Lodge, 2006). The principal rejoinder contends that these patterns can be reconciled with a Bayesian model of rational learning. People with different prior beliefs may update in divergent ways not due to motivation but because their priors generate different posterior beliefs about the same newly learned information (e.g., Gerber and Green, 1999; Bullock, 2009). A slew of recent studies further contend that resistance is not the modal way in which people respond to political information. For example, people seem to consistently update beliefs in the direction of fact checks, irrespective of their priors, across settings and across

time (Walter et al., 2020; Porter, Velez and Wood, 2023; Coppock et al., 2023).²

Neither critique, however, addresses the puzzle animating this paper. Gerber and Green (1999) emphasize that the Bayesian learning model is concerned with beliefs, “which measure a voter’s assessments of objective characteristics of the political world” (193), but not evaluations or preferences. As a result, the Bayesian learning debate offers little insight into why belief and attitude change might diverge. Moreover, fact checks that successfully correct false beliefs rarely have discernible effects on downstream attitudes or behaviors, even in well-powered studies (e.g., Nyhan et al., 2020; Porter, Velez and Wood, 2022).

The literature therefore offers little consensus about how to interpret cases in which information fails to shape attitudes despite altering beliefs. A motivated reasoning account would suggest that people are more intensely motivated to preserve attitudes than to preserve beliefs and therefore may rationalize away the connection between newly learned facts and their prior opinions. Studies that draw upon motivated reasoning have concluded, for instance, that “beliefs are not so much building blocks of attitudes but consequences of attitudes” (Lawrence and Sides, 2014, 2). Other studies take a more conservative approach, drawing conclusions about domain-specific attitudes but not about persuasion in general.³

In this paper we caution against drawing broad conclusions about persuasion from these kinds of experimental results, especially before belief relevance is taken into account. We begin by defining the difference between beliefs and attitudes and specifying a function that links belief to attitude change. Doing so helps clarify why, although the Bayesian versus motivated reasoning debate does not address our puzzle, existing intuitions about information processing can be incorporated within a broader model of attitude change. Drawing

²This second rejoinder has two caveats. First, whether this sort of evidence can help distinguish motivated reasoning from Bayesian learning is subject to continued debate within the modeling literature (Gerber and Green, 1999; Little, 2025). Second, so-called “parallel updating” may be limited to cases in which information is presented in a journalistic, tonally neutral fashion without group cues (Coppock, 2023).

³An example is Dias et al. (2024), who find that correcting out-party misperceptions does not produce robust downstream effects on Americans’ partisan animosity or support for democratic norms. The authors conclude that out-party perceptions are simply not at the crux of partisan animosity, especially as these perceptions are unstable over time and appear weakly held.

on the expectancy value (EV) model, we argue that beliefs may take on different weights in this attitude function due to varying relevance to an attitude. Finally, we argue that typical designs ignore individual-level variation in belief relevance, contributing to the conclusion that attitudes do not respond to belief change.

Distinguishing Beliefs From Attitudes

While studies of persuasion frequently measure relevant beliefs and attitudes in tandem,⁴ the distinction is key. Whereas beliefs pertain to perceptions of facts or objective states of the world, attitudes lack inherent truth value, instead representing personal preferences, values, or normative judgments about what ought to be (Petty, 2024, 1276). One might hold the belief that a candidate is progressive (a factual claim), while viewing that characteristic negatively (an evaluative position).

According to Bayesian models, individuals possess prior beliefs about a state of the world and update these beliefs when presented with new information, weighing both the strength of their prior beliefs and the credibility of the new evidence. In the simplest model, the posterior probability that one’s view of a state of the world, b , is true given some evidence, e , depends on the prior probability of b (i.e., $\Pr(b)$) and the likelihood of e given b :

$$\Pr(b \mid e) = \frac{\Pr(e \mid b) \Pr(b)}{\Pr(e \mid b) \Pr(b) + \Pr(e \mid \neg b) \Pr(\neg b)}$$

Attitudes, as evaluative judgments, do not hold truth value *per se*, but they often rest on a foundation of beliefs or considerations (Fishbein, 1979). Specifying a *belief-attitude function* enables us to incorporate Bayesian learning within a model of attitude change. Consider an individual forming an attitude A toward an object (e.g., a policy proposal). Suppose this person holds a set of n beliefs, b_i , about various attributes of the policy, such as whether the policy is effective, necessary, or costly. Each belief b_i can be represented as

⁴For a review of fact-checking studies in which this design is common, see Porter and Wood (2024a).

the perceived probability that the policy possesses attribute i . Using the Bayesian model, $\Pr(b_i | e)$ can be updated in the light of new evidence e . However, to translate those updated beliefs into an overall evaluative judgment, an additional function is necessary. Here, $f(\cdot)$ is an aggregation rule that maps the updated belief probabilities into a summary evaluation:

$$A = f(\Pr(b_1), \Pr(b_2), \dots, \Pr(b_n)).$$

The implications of a belief can vary substantially across individuals. One reason is that people may differ in their evaluation of the attribute implied by b_i . Those who believe a candidate is progressive may hold positive, negative, or neutral affect toward progressives, changing the implications of those beliefs. This is the intuition behind psychology's expectancy value (EV) model (Fishbein and Ajzen, 1975, 2010), one of several classic frameworks that conceive of attitudes as an aggregation of beliefs and evaluations.⁵ A second reason is that politicians or media figures often shape the considerations citizens bring to bear on their political opinions through issue framing. In summative models like EV, this corresponds to shifting the weights associated with beliefs (Jerit, 2009; Chong and Druckman, 2007), though the precise mechanism behind changes in weights can vary (Nelson, Oxley and Clawson, 1997; Slothuus, 2008). A frame may increase the weight of a consideration either by increasing its cognitive accessibility or by changing the explicit importance we consciously assign it (Nelson and Oxley, 1999). Formally, this can be expressed as:

$$A = f(w_1 \cdot \Pr(b_1), w_2 \cdot \Pr(b_2), \dots, w_n \cdot \Pr(b_n))$$

where w_i represents the weight assigned to each belief.

In short, the extent to which changes in beliefs result in shifts in attitudes depends fundamentally on the relevance of those beliefs to individuals' judgments. Beliefs deemed

⁵In the EV model, an overall attitude toward an object is produced by summing across all relevant attributes, indexed by i , in a process modeled symbolically as $A \propto \sum e_i \Pr(b_i)$. e_i is the attribute's evaluation.

irrelevant or peripheral are less likely to result in attitude change, even if convincingly corrected. Here, we use *relevance* or *importance* to signify the weights on beliefs, agnostic about the extent to which these weights capture cognitive salience, a conscious judgment of *perceived* relevance, or some mix.

The Role of Relevance

To illustrate why understanding the variable attitudinal effects of information requires accounting for relevance, consider attitudes toward immigrants. Motivated by theories of power threat (e.g., Key, 1949; Oliver and Mendelberg, 2000), many have hypothesized that immigration attitudes hinge upon beliefs about the size of the immigrant population, and therefore that correcting persistent overestimates of immigrants' numeracy should reduce hostility (Citrin and Sides, 2008; Wong, 2007). A substantial body of evidence now suggests that correcting misperceptions about numeracy is ineffective at altering anti-immigrant attitudes, leading a recent paper to conclude that misperceptions may be the consequence rather than the cause of hostility toward immigrants (Hopkins, Sides and Citrin, 2019). More broadly, and more worryingly, the evidence could suggest that "correct information can change factual perceptions but not necessarily related attitudes" (316).

This interpretation coheres with theories of motivated reasoning, but it assumes that people heavily weight immigrants' numeracy when formulating immigration attitudes. Recent studies highlight alternative beliefs that, when corrected, produce stronger effects on these judgments. Grigorieff, Roth and Ubfal (2020) offer evidence that whereas a bundle of information about immigrants' characteristics and numeracy alters immigration policy preferences, a lighter treatment containing only information about numeracy does not. Likewise, experimental interventions that supply knowledge about immigration's impact on crime, jobs, and taxes (Abascal, Huang and Tran, 2021) and about the difficulty of the legal immigration process (Kustov and Landgrave, 2025) have substantially affected immigration attitudes, suggesting that the limited downstream effects of past interventions may be "the

result of not providing relevant information” (Kustov and Landgrave, 2025, 1). In Appendix B, we present evidence that respondents who favor restricting immigration are indeed far more likely to justify their attitude using beliefs about immigration’s effects on crime, public services, or economic opportunity than using arguments about numeracy.

Focusing on the potential role of belief relevance helps reconcile some of the seemingly contradictory findings in the persuasion literature. Though there is ample evidence that factual corrections reliably improve belief accuracy with little risk of backfire, evidence of attitudinal shifts has been decidedly more mixed. If fact-checking studies target beliefs that are salient or viral but largely distal to the formation of relevant attitudes for much of the population, this could help explain the consistent observation of belief change without corresponding shifts in attitudes.

Challenges in Identifying Belief Relevance Weights

Researchers use various strategies to identify beliefs for correction, each with distinct theoretical implications. Misinformation scholars typically target beliefs based on their prevalence, drawing on fact-checking databases to identify widely circulating false claims (see Porter, Velez and Wood, 2023). This approach prioritizes reach over psychological centrality. Other researchers rely on media analysis or researcher intuition about common misperceptions, potentially missing beliefs that are crucial but less visible in public discourse. A more systematic approach, common in attitude research but rare in misinformation studies, uses pilot surveys to identify “modal salient beliefs” within a population before designing interventions.

Population-level approaches to identifying influential beliefs—whether based on expert judgment, viral spread, or aggregate network patterns—may miss crucial individual variation in belief relevance. For instance, while concerns about immigration’s fiscal impact may strongly predict immigration attitudes in aggregate data, this belief may be peripheral or irrelevant for many individuals. Even sophisticated network methods that map belief-

attitude connections typically do so at the group level (e.g., Fishman and Davis, 2022), potentially obscuring individual-level heterogeneity in belief centrality (Brandt and Morgan, 2022). Moreover, recent studies that attempt to construct belief networks at the individual level assess relationships between political issues determined *a priori*, which may obscure idiosyncratic connections between beliefs and attitudes (Brandt, 2022). For example, a belief about “universal basic income” may be central to an individual’s views about politics, and failure to include relevant survey questions about the issue will produce only a partially observable network.

We propose an alternative method: using large language models (LLMs) to identify personally salient beliefs through direct dialogue with participants. Rather than assuming beliefs relevant at the population level will matter for all individuals, this approach elicits each participant’s own stated reasons for holding their attitudes. By engaging participants in an open-ended discussion about their views, we can identify the specific beliefs they spontaneously invoked to justify their positions—beliefs that are, on the respondent’s own terms, identified as relevant to their individual attitude structures. Upon identifying important attitude-relevant beliefs, we construct counterarguments targeting those beliefs to assess how this affects downstream attitudes.

Using our design, we elicit issue attitudes (e.g., supporting DACA, opposing restrictions on automatic firearms) and focal justifications (e.g., improving the educational attainment of all Americans, protecting Second Amendment rights) for each individual in our study, which we use to produce tailored measures of attitude strength and belief strength (see Appendix A for question wording). Moreover, we instruct an LLM—OpenAI’s GPT-4o—to create personalized corrective messages that target those focal justifications (hereafter, focal beliefs) and more distal justifications reflecting beliefs not disclosed in the open-ended dialogue that nonetheless support the issue attitude (e.g., economic contributions of undocumented immigrants and their children).⁶ This allows us to directly test whether heavily

⁶Specifically, we used the GPT-4o-2024-08-06 model for generating counterarguments and GPT-4o-mini

weighted considerations produce more attitude change than those that are less relevant to the participant's attitude.

Our primary hypotheses build on prior work in factual corrections, motivated reasoning, and attitude formation. We propose the following sequential expectations:

Assessing the “First Stage” of Belief Change

Hypothesis 1a (Parallel Updating). Studies of factual corrections have increasingly found that people consistently update their beliefs in the direction of evidence, irrespective of their prior beliefs (e.g., Porter, Velez and Wood, 2022; Coppock, 2023). These studies generate an expectation that *focal* and *distal* belief counterarguments will reduce focal and distal belief strength (respectively) relative to control conditions, and to similar degrees.

Individuals presented with a factual argument (either focal or distal) will update their beliefs toward the corrective information to similar degrees.

Hypothesis 1b (Differential Updating). Drawing on theories of motivated reasoning, individuals may process factual arguments differently depending on how central the challenged beliefs are to their broader belief system. While a strong form of motivated reasoning suggests complete resistance or backfire effects (Kunda, 1990; Taber and Lodge, 2006), people may also act as “cautious Bayesians” who update their beliefs to varying degrees based on the centrality and strength of their prior beliefs (Hill, 2017). Under this account:

The magnitude of belief updating will vary based on belief relevance, with focal beliefs showing either no movement or significantly smaller updates compared to distal beliefs. Specifically: (i) For focal beliefs, individuals will either maintain their prior beliefs or exhibit minimal updating in response to factual arguments; (ii) For distal beliefs, which are less central to attitudes, individuals will show

for carrying out the semi-structured chat.

larger shifts toward the corrective information.

Assessing the Belief-Attitude Function

Hypothesis 2a (Equal Effectiveness). It may be that *any* counterargument—whether targeting a focal or distal belief—has a comparable effect on attitudes, provided the individual updates and sees the corrected information as relevant enough to reshape their broader evaluation. This represents what seems to be the operating assumption within the literature.

Counterarguments to focal and distal beliefs will be equally effective at shifting downstream attitudes.

Hypothesis 2b (Focal Belief Dominance). Alternatively, if focal beliefs are highly weighted in the formation of attitudes, changing these beliefs should be more influential in shifting attitudes than changing more peripheral, distal beliefs.

Factual arguments against focal beliefs (i.e., those identified by participants as central to their opinions) will produce greater changes in downstream attitudes than counterarguments for distal beliefs.

Research Design

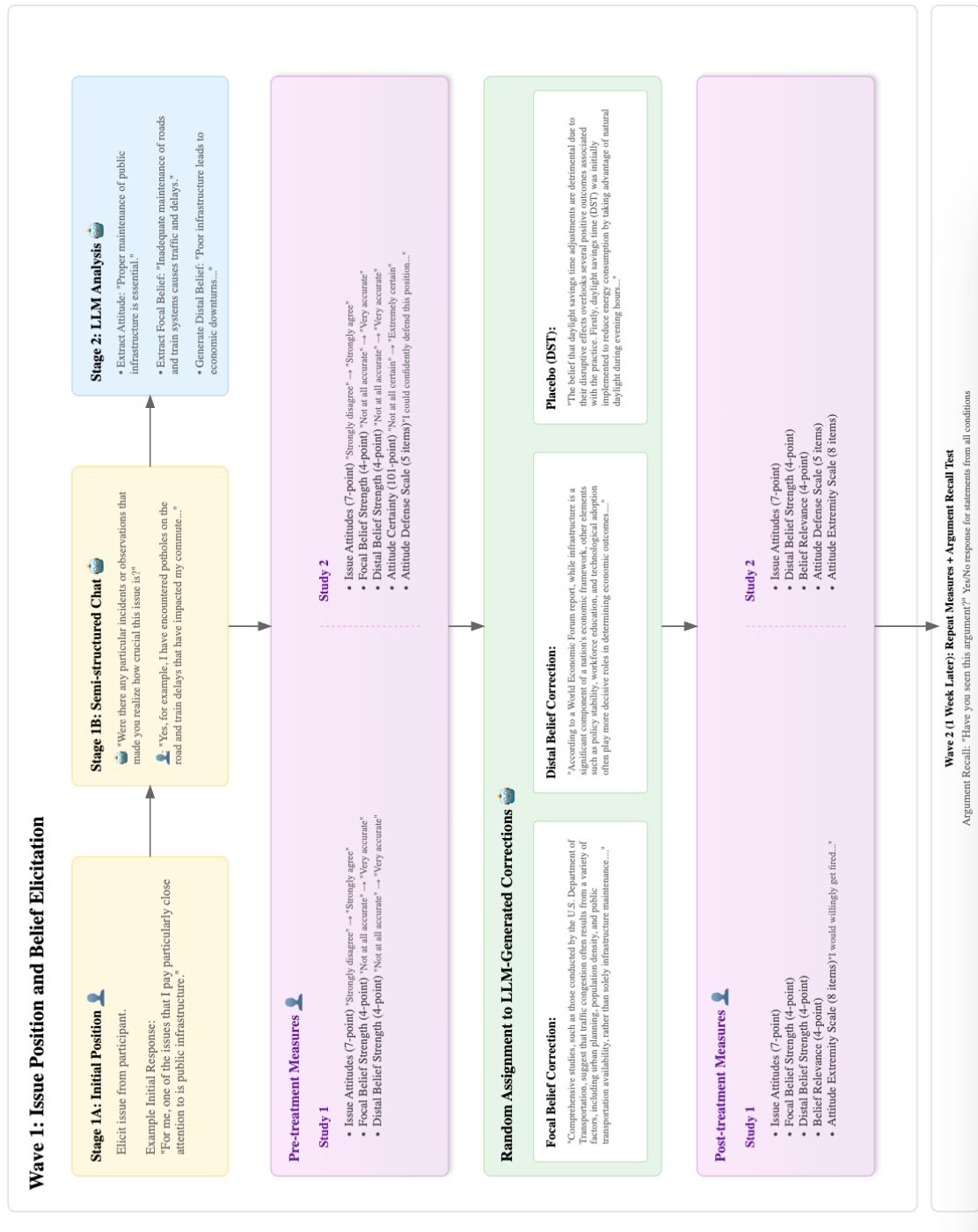
Our experimental design uses LLMs at various stages of the process to recover key theoretical quantities and generate experimental stimuli (see Figure 1 for a diagram). Text classification, summarization, and generation are among the strengths of LLMs (Mellon et al., 2022), and previous work has shown that they can be productively integrated into surveys to measure deeply held beliefs and attitudes (Velez and Liu, 2024) and produce compelling arguments (Bai et al., 2023). We leverage these strengths to (1) elicit issue attitudes; (2) capture justifications for those attitudes (“focal beliefs”); and (3) generate counterarguments that randomly target those “focal beliefs” versus unmentioned, but potentially relevant, beliefs

that might support those attitudes (“distal beliefs”).

We carry out a pre-post experimental design that comprises three stages (Clifford, Sheagley and Piston, 2021). First, we elicit issue positions from participants. In Study 1, we randomly assigned participants to either a closed-ended slate of issue positions or open-ended question asking them to report a deeply held issue position. In Study 2, we solely rely on the latter question. Second, in both studies, we have all participants engage in a semi-structured discussion about their issue position and their justifications for holding the issue position. Once the conversation is complete, these data are passed to an LLM that summarizes the participant’s issue attitude, summarizes their primary justification for holding that attitude (“focal belief”), generates a related, but unmentioned, justification for holding the attitude (“distal belief”), and produces three factual argument vignettes: (1) a vignette countering the “focal belief”; (2) a vignette countering the “distal belief”; and (3) a placebo vignette producing an argument about Daylight Savings Time. Participants were informed of their interactions with AI models in the consent form and debriefing materials.

Following this phase, we measure issue attitudes on a seven-point Likert item ranging from “strongly disagree” to “strongly agree,” focal belief strength on a four-point scale ranging from “not at all accurate” to “very accurate,” and distal belief strength using the same scale. In Study 2, we also measure pre-treatment scores on an “attitude defense” scale developed in Velez and Liu (2024). After the pre-treatment measurement step, participants are randomly assigned to see one of the tailored vignettes, with participants equally likely to be assigned to a “focal belief counterargument,” “distal belief counterargument,” or “placebo” condition. These vignettes involved approximately paragraph-long arguments disputing the assigned belief. The LLM was instructed to incorporate logical arguments and data when available.

In the post-treatment phase, we measured attitude strength and belief strength for the focal and distal beliefs. In Study 1, we additionally measured a multi-item “attitude ex-



Focal Belief Correction:

"Comprehensive studies, such as those conducted by the U.S. Department of Transportation, suggest that traffic congestion often results from a variety of factors, including urban planning, population density, and public transportation availability, rather than solely infrastructure maintenance..."

Distal Belief Correction:

"According to a World Economic Forum report, while infrastructure is a significant component of a nation's economic framework, other elements such as policy stability, workforce education, and technological adoption often play more decisive roles in determining economic outcomes..."

Placebo (DST):

"The belief that daylight savings time adjustments are detrimental due to their disruptive effects overlooks several positive outcomes associated with the practice. Firstly, daylight savings time (DST) was initially implemented to reduce energy consumption by taking advantage of natural daylight during evening hours..."

Figure 1: Overview of the Two-Wave Design and Experimental Procedures. In *Wave 1*, participants first provide an issue they care deeply about (*Stage 1A*), then engage in a semi-structured dialogue with an LLM to clarify their justifications (*Stage 1B*). The LLM identifies participants' "focal" beliefs, generates "distal" beliefs, and constructs tailored counterarguments (*Stage 2*). Pre-treatment measures are taken and participants are then randomly assigned to receive one of three messages (focal belief counterargument, distal belief counterargument, or placebo) before completing post-treatment measures. In *Wave 2*, approximately one week later, participants repeat the attitude and belief measures to assess the durability of belief and attitude change.

tremity” scale developed in Velez and Liu (2024), whereas in Study 2, we measured “attitude extremity,” “attitude defense,” and certainty on a 101-point scale. In addition to attitudinal measures, we asked participants to report the relevance of their beliefs to their attitudes on a four-point scale ranging from “not at all important” to “very important” in Studies 1 and 2. In Study 2, we provided participants with a text entry box to indicate “omitted beliefs” that support their attitude, and a rating scale. The latter measure was included to capture whether participants selectively bolstered unmentioned beliefs as a defense mechanism in the face of corrective information.

In both studies, we carry out a second wave approximately one week after the first wave. In this study, we measure the full set of beliefs and attitudes measured in the initial wave, along with measures of argument recall. Specifically, we provide participants with two randomly selected sentences from the three conditions (i.e., “focal belief counterargument,” “distal belief counterargument,” and placebo) and ask them to identify the sentences that they recall seeing in the previous study. This serves as a test of selective memory, such that participants may forget corrective information that bears on their deeply held beliefs or attitudes.

Our first study was conducted on CloudResearch Connect ($n=2,494$). Wave 1 was fielded through November 1-5, 2024, and wave 2 ($n = 2,096$) through November 15-22. Study 2 sought to replicate results with a sample from Prolific. Wave 1 ($n = 1,913$) was fielded through November 26-27, 2024, and wave 2 ($n = 1,712$) through December 5-12. CloudResearch Connect and Prolific are both high-quality sample providers that have been used extensively in social science research (Stagnaro et al., 2024).

Can We Recover Strong Attitudes and Focal Beliefs?

We focus on identifying participants’ deeply held attitudes. Though this may decrease our ability to detect belief and attitude change, we do so for two important reasons. Relative to

“non-attitudes” or those that are not very salient to participants, deeply held attitudes ought to have a richer set of connected considerations (Zaller, 1992). Respondents will be more likely to call to mind relevant beliefs in conversation, enabling us to observe how intervening on a belief affects attitudes. Second, focusing on deeply held attitudes provides a “harder test” of our theoretical predictions, as any observed effects are likely to represent meaningful attitude change rather than temporary shifts in superficial opinions.

To what extent can we recover deeply held attitudes? As shown in Table 1, we observe attitude strength scores of 4.85 (SE = .01) in Study 1 and 4.84 (SE = .02) in Study 2. Given that attitude strength is measured on a five-point scale, this means that the vast majority of participants select the maximum score, indicating strong attitudes in the aggregate. Indeed, we find that 89% of participants in Study 1 select the maximum score on the Likert item, whereas 88% do so in Study 2. This dovetails with Velez and Liu (2024) who find that LLMs can successfully summarize issue positions and recover deeply held attitudes that persist over time.

Next, we turn to belief strength. Our theoretical tests hinge on the capability of LLMs to properly summarize focal justifications for attitudes. In Study 1, we find a mean belief strength score of 3.64 (SE = .01), with 72% of participants selecting the maximum score; Study 2 replicates this general pattern, with a mean belief strength score of 3.67 (SE = .02) and 72% of participants selecting the maximum score. Directly comparing focal beliefs to distal beliefs, we observe a difference in belief strength of .32 scale points in Study 1 and .31 scale points in Study 2, suggesting that the focal beliefs retrieved are stronger than distal beliefs.

Additionally, we can assess whether participants see the focal beliefs as relevant to their attitudes. In Study 1, we observe a mean belief relevance score of 3.72 with 76% selecting the maximum option. In Study 2, the mean belief relevance score is 3.84 with 85% selecting the maximum. This can be compared to the distal belief, where scores are

approximately .27 scale points lower in Study 1 and .23 in Study 2. Thus, focal beliefs are generally seen as relevant to attitudes, and *more* relevant than distal beliefs.

Table 1: Summary Statistics for Variables Across Studies.

Variable	Study 1	Study 2
Attitude Strength	4.85 (0.01); 89.4% selected max	4.84 (0.02); 88.4% selected max
Belief Strength (Focal)	3.64 (0.01); 71.9% selected max	3.67 (0.02); 71.6% selected max
Belief Strength (Distal)	3.32 (0.02); 55.3% selected max	3.36 (0.03); 57.3% selected max
Belief Relevance (Focal)	3.72 (0.01); 75.7% selected max	3.84 (0.02); 84.6% selected max
Belief Relevance (Distal)	3.45 (0.02); 60.1% selected max	3.61 (0.03); 70.2% selected max

Note: Mean (SE) and Percentage Selecting Maximum Option.

Overall, the evidence suggests that we are successfully recovering deeply held attitudes and relevant beliefs. Moreover, the strong performance of the distal beliefs suggests that the LLM is also producing plausible distal beliefs that may support participants' attitudes, even if they are not as central to their reasoning. This provides reassuring evidence that our approach successfully identifies strong attitudes and focal beliefs.

Inducing Belief Change

We now examine whether and how these beliefs respond to corrective information. This analysis provides a possible test of competing expectations about belief change. According to the differential updating hypothesis (H1b), we should observe resistance or even backfire effects when correcting focal beliefs that participants use to justify their attitudes, whereas a parallel updating process (H1a) might hold that counterarguments have similar efficacy regardless of belief strength.

To assess the effects of arguments on focal and distal belief strength, we estimate OLS regression models with HC2 standard errors. The models take the following form:

$$Y_i = \beta_0 + \beta_1 \text{Focal Belief Counterarg}_i + \beta_2 \text{Distal Belief Counterarg}_i + \mathbf{X}_i' \boldsymbol{\gamma} + \epsilon_i \quad (1)$$

where Y_i is post-treatment belief strength for participant i , $\mathbf{X}_i$ is a vector of pre-treatment covariates, and the omitted category is the placebo condition. The coefficients of interest are β_1 and β_2 , which capture the average treatment effects of focal and distal belief arguments relative to the placebo. In Study 1, we adjust for pre-treatment attitude strength, focal belief strength, and distal belief strength. In Study 2, we adjust for the covariates, as well as pre-treatment scores on a multi-item “attitude defense” scale.

We directly compare effect estimates across two outcomes: focal belief strength and distal belief strength. For each outcome, we estimate separate regression models where the dependent variable is either the focal or distal belief strength. This allows us to test our competing hypotheses in several ways:

First, to assess H1a, we compare β_1 when focal belief strength is the outcome to β_2 when distal belief strength is the outcome. Similar effect sizes would indicate that factual arguments are equally effective regardless of belief type—that is, focal belief arguments move focal beliefs just as much as distal belief arguments. Second, for H1b, we examine whether focal beliefs are more resistant to change than distal beliefs. Under this hypothesis, β_2 might be negative when distal belief strength is the outcome (indicating successful belief change), but β_1 could be positive or null when focal belief strength is the outcome (indicating resistance or backfire). Alternatively, β_1 could be negative but smaller in absolute magnitude than β_2 , consistent with a “cautious Bayesian” view.

Turning to Table 2, we find evidence that is more consistent with H1a. For example, in the first wave of Study 1, we find that focal belief arguments reduce focal belief strength by .32 scale points and distal belief strength by .13 scale points relative to control (SE = .03; $p < .001$). Turning to distal belief strength, we observe similar effect sizes for the distal argument, with reductions in distal belief strength of .31 scale points (SE = .03; $p < .001$) and .12 scale points (SE = .03; $p < .001$) in focal belief strength. In the first wave of Study 2, we find comparable effect sizes, with focal belief counterarguments reducing focal

Table 2: Effect of Interventions on Belief Strength (Wave 1)

	Study 1	Study 1	Study 2	Study 2
	Focal Belief	Distal Belief	Focal Belief	Distal Belief
Focal Correction	−0.32*** (0.03)	−0.13*** (0.03)	−0.32*** (0.03)	−0.17*** (0.03)
Distal Correction	−0.12*** (0.02)	−0.31*** (0.03)	−0.12*** (0.03)	−0.28*** (0.03)
Num.Obs.	2477	2475	1898	1897
R2	0.406	0.587	0.417	0.544
Covariates Included	✓	✓	✓	✓

+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

belief strength by .32 scale points (SE = .03; $p < .001$) and distal belief arguments reducing distal belief strength by .28 scale points (SE = .03; $p < .001$). Thus, focal arguments are approximately just as effective in moving focal beliefs as distal arguments are in moving distal beliefs.

Importantly, these effects persist into the second wave of each study, approximately one week after the initial exposure. In Study 1, focal belief counterarguments continue to reduce focal belief strength by .24 scale points (SE = .04; $p < .001$). Similarly, distal belief arguments maintain their impact on distal belief strength though the effect is somewhat smaller in magnitude ($\beta = -.15$, SE = .04; $p < .001$). Study 2 replicates this pattern of persistence: focal belief arguments continue to reduce focal belief strength by .26 scale points (SE = .04; $p < .001$), while distal belief arguments maintain their effect on distal belief strength ($\beta = -.23$, SE = .04; $p < .001$). The durability of these effects across both studies suggests that counterattitudinal arguments produce lasting changes in belief strength, rather than reflecting momentary shifts arising due to other processes such as demand effects.

The similarity in effect sizes between focal and distal belief counterarguments suggests that both types of beliefs are movable and that arguments targeting either type of belief can have spillover effects on the other, which is inconsistent with H1b. Moreover, our finding that focal argument effects exhibit more persistence than distal argument effects suggests

Table 3: Effect of Interventions on Belief Strength (Wave 2)

	Study 1	Study 1	Study 2	Study 2
	Focal Belief	Distal Belief	Focal Belief	Distal Belief
Focal Correction	−0.24*** (0.04)	−0.10** (0.04)	−0.26*** (0.04)	−0.13** (0.04)
Distal Correction	−0.07* (0.03)	−0.15*** (0.04)	−0.08* (0.03)	−0.23*** (0.04)
Num.Obs.	2040	2038	1679	1679
R2	0.198	0.343	0.264	0.355
Covariates Included	✓	✓	✓	✓

+ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

that interventions targeting beliefs central to attitudes produce *more* durable belief change than those targeting peripheral beliefs—perhaps because they are more relevant to deeply important attitudes. This provides additional evidence against motivational accounts that predict greater resistance to updating focal beliefs, while also underscoring an important pattern: the apparent equivalence between focal and distal belief arguments may only hold in the short term.

Inducing Attitude Change

Demonstrating the effectiveness of factual arguments on beliefs, we assess whether those targeting focal beliefs exhibit larger effects on attitudes relative to those targeting distal beliefs. We review the findings of each test here before turning to a comprehensive meta-analytic test of the difference between focal and distal arguments. In the first wave of Study 1, the focal counterargument reduces attitude strength by 0.12 scale points ($SE = 0.02$, $p < 0.001$), while the distal argument has approximately half the effect, decreasing strength by 0.06 scale points ($SE = 0.02$, $p < 0.01$). Turning to attitude extremity, we find a .05 scale point reduction ($SE = .07$; $p = \text{n.s.}$) in scores for the focal argument and .01 scale point increase for the distal argument ($SE = .07$; $p = \text{n.s.}$). Though neither are significant, the focal argument moves estimates in the expected direction. In the first wave of Study 2, we

Table 4: Effect of Interventions on Attitudes (Wave 1)

	Study 1 Strength	Study 1 Extremity	Study 2 Strength	Study 2 Extremity	Study 2 Certainty	Study 2 Defense
Focal Correction	-0.12*** (0.02)	-0.05 (0.07)	-0.13*** (0.03)	-0.06+ (0.04)	-3.73*** (0.68)	-0.13** (0.05)
Distal Correction	-0.06** (0.02)	0.01 (0.07)	-0.09*** (0.03)	0.001 (0.04)	-2.16*** (0.65)	-0.09+ (0.05)
Num.Obs.	2476	2479	1900	1902	1897	1901
R2	0.367	0.026	0.271	0.117	0.279	0.695
Covariates Included	✓	✓	✓	✓	✓	✓

+ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

largely replicate the findings from Study 1. The focal argument reduces attitude strength by .13 scale points ($SE = .03$; $p < .001$) relative to the distal argument, which produces an effect estimate of .09 scale points ($SE = .03$; $p < .001$). Turning to attitude extremity, we find modest evidence that focal arguments reduce extremity by .06 scale points ($SE = .04$, $p < .10$), while distal arguments have no discernible effect with estimates hovering around zero ($SE = .04$, $p = n.s.$). Focusing on our additional attitudinal measures, we detect a -3.7 scale point reduction on a 101-point attitude certainty scale ($SE = .68$; $p < .001$) among those exposed to the focal argument, whereas this estimate is -2.16 ($SE = .65$; $p < .001$) among those exposed to the distal argument. Finally, we observe a -.13 scale point reduction in “attitude defense” ($SE = .06$; $p < .01$) among those exposed to the focal argument, which is .04 scale points larger than the effect we observe for the distal argument ($\beta_2 = -.09$; $SE = .05$; $p < .10$).

Overall, effect sizes tend to be small, reflecting the choice to target deeply held attitudes. Most importantly, however, we find consistent evidence that focal belief counterarguments produce larger immediate attitude change than distal belief counterarguments across multiple measures and studies.

Turning now to our second wave, we find that effects on attitudes generally weaken. In Study 1, while the immediate effects on attitude strength dissipate ($\beta = -.02$, $SE =$

Table 5: Effect of Interventions on Attitudes (Wave 2)

	Study 1 Strength	Study 1 Extremity	Study 2 Strength	Study 2 Extremity	Study 2 Certainty	Study 2 Defense
Focal Correction	-0.02 (0.03)	-0.09 (0.07)	-0.06 (0.04)	-0.14+ (0.08)	-2.57*** (0.71)	-0.10 (0.07)
Distal Correction	0.03 (0.03)	-0.10 (0.07)	0.02 (0.03)	-0.003 (0.08)	-1.00 (0.72)	-0.04 (0.07)
Num.Obs.	2043	2042	1687	1680	1678	1680
R2	0.103	0.027	0.112	0.079	0.241	0.447
Covariates Included	✓	✓	✓	✓	✓	✓

+ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

.03, $p = \text{n.s.}$ for focal arguments; $\beta = .03$, $\text{SE} = .03$, $p = \text{n.s.}$ for distal arguments), we continue to observe modest reductions in attitude extremity, though these effects do not reach conventional levels of statistical significance ($\beta = -.10$, $\text{SE} = .07$, $p = \text{n.s.}$ for focal arguments; $\beta = -.11$, $\text{SE} = .07$, $p = \text{n.s.}$ for distal arguments). Study 2 reveals a similar pattern of attenuation, but with some notable exceptions. While effects of the focal argument on attitude strength and extremity weaken ($\beta_1 = -.06$, $\text{SE} = .04$, $p = \text{n.s.}$ for attitude strength; $\beta_1 = -.14$, $\text{SE} = .08$, $p < .10$ for extremity), focal arguments continue to reduce attitude certainty by 2.57 points on our 101-point scale ($\text{SE} = .71$, $p < .001$). In contrast, the effect of distal belief arguments on certainty is no longer significant ($\beta_2 = -1.00$, $\text{SE} = .72$, $p = \text{n.s.}$). Effects on attitude defense show a similar pattern of persistence for focal arguments ($\beta_1 = -.10$, $\text{SE} = .07$, $p = \text{n.s.}$) compared to distal arguments ($\beta_2 = -.04$, $\text{SE} = .07$, $p = \text{n.s.}$), though neither reaches conventional levels of statistical significance. Taken together, we observe mixed evidence of persistence, but generally observe larger effect estimates in Wave 2 for focal counterarguments relative to distal counterarguments.

Assessing Differences in Attitude Change

We now conduct an exploratory analysis to formally assess differences between focal and distal counterarguments on attitudes. While focal arguments generally produce larger effect

estimates, examining each attitudinal measure separately may obscure the overall pattern of effects and reduce our ability to detect meaningful differences between argument types. A meta-analytic approach allows us to systematically compare the relative effectiveness of focal versus distal arguments while accounting for both within- and between-measure heterogeneity in effect sizes.

We employ a random effects meta-analysis for several reasons. First, our attitudinal measures—though related—capture distinct theoretical constructs ranging from strength to extremity to defense. We should therefore expect some variation in the true effect sizes across measures. Second, these measures operate on different scales (e.g., 4-point versus 101-point), which could produce mechanical differences in effect magnitudes. Third, we collected these measures across two independent studies and waves, introducing another potential source of effect heterogeneity.

Pooling across all attitudinal measures, we find that focal arguments produce significantly larger effects than distal belief arguments. In 11 out of 12 comparisons, we observe larger effect sizes for the focal argument relative to the distal argument. The average difference in treatment effects is $-.06$ control-group standard deviation units ($SE = .02$, $p < .001$). We fail to reject the null hypothesis of effect homogeneity ($Q(11) = 8.79$, $p = .64$), suggesting that the effect of focal arguments is comparable across different attitudinal measures, waves, and studies. Looking at individual comparisons, the largest differences appear in attitude certainty, where focal arguments outperform distal arguments by roughly $.10$ standard deviation units in both Wave 1 ($d = -.104$, $SE = .050$) and Wave 2 ($d = -.105$, $SE = .052$) of Study 2. The consistency of positive differences between the focal and distal arguments provides support for H2b (Focal Belief Dominance) while suggesting that larger attitudinal effects due to targeting focal beliefs generalizes across different measures.

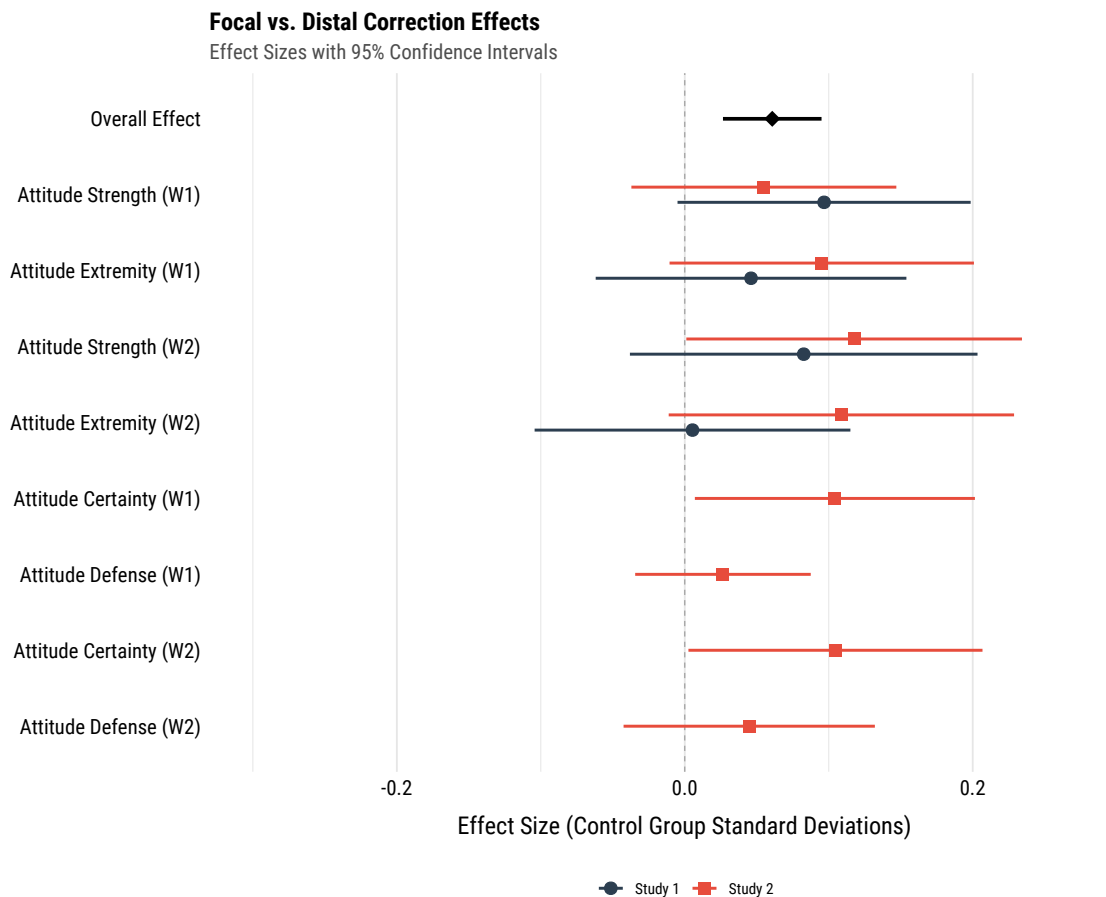


Figure 2: Results from a random effects meta-analysis with correlated sampling errors showing mean differences between the focal and distal argument conditions across two studies. Points show the estimated effects and horizontal lines represent 95% confidence intervals. The diamond shows the overall random effects meta-analytic estimate.

Mechanisms of Attitude Change: Relevance and Recall

Belief relevance In addition to belief strength, we consider auxiliary outcomes that help clarify why focal arguments generally have larger and more persistent effects than distal arguments. First, we focus on belief relevance, which is measured on a four-point scale capturing the relevance of focal and distal beliefs to attitudes. We regress belief relevance for the two relevance outcomes on treatment indicators and pre-treatment covariates with HC2 robust standard errors.

Turning to Table 6, we find evidence that participants decrease belief relevance when exposed to the corresponding argument. In Wave 1 of Study 1, focal arguments decrease focal belief relevance by -.167 scale points ($SE = .025$; $p < .001$) and distal arguments decrease distal belief relevance by -.205 ($SE = .032$; $p < .001$), with some evidence of spillover on focal belief relevance ($\hat{ATE} = -.120$; $SE = .030$; $p < .001$). We replicate these findings in Study 2, with reductions of .09 (.13) scale points in focal (distal) relevance among those assigned to the focal (distal) argument.

On their face, these findings are consistent with a “bolstering” process (Jervis, 2006, 653), where once salient considerations are supplanted by alternative considerations, leading individuals to discount their previous reliance on these beliefs when reporting their attitudes. However, it is worth noting that the self-reports of relevance and belief strength are intertwined, with moderate correlations between the two.⁷ This may be because beliefs that are held with less certainty are perceived as less relevant to an attitude.

To directly assess the bolstering mechanism, we included a text entry option in Study 2 when asking about belief relevance. In addition to presenting them with focal and distal beliefs, we told participants “if there are more important reasons why you hold this view that are not listed, please tell us using the *Other* text box.” If participants are simply

⁷In the initial wave of Studies 1 and 2, the correlation between focal belief strength and self-reported relevance is .40 and .46, respectively. The correlation between distal strength and relevance is .59 and .58, respectively.

Table 6: Effect of Interventions on Belief Relevance

	Study 1 Focal Relevance	Study 1 Distal Relevance	Study 2 Focal Relevance	Study 2 Distal Relevance
Focal Correction	−0.17*** (0.03)	−0.12*** (0.03)	−0.09** (0.03)	−0.06+ (0.04)
Distal Correction	−0.04+ (0.02)	−0.20*** (0.03)	0.005 (0.03)	−0.13*** (0.04)
Num.Obs.	2470	2471	1679	1679
R2	0.202	0.367	0.160	0.227
Covariates Included	✓	✓	✓	✓

+ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

substituting a corrected belief for another, we ought to see additional considerations being offered in the counterargument conditions compared to the control condition. This would provide direct evidence of the theorized process whereby individuals compensate for weakened beliefs by elevating the importance of alternative considerations. However, turning to Table 7, we fail to find evidence that this is what participants are doing, with small estimates hovering around zero for both experimental conditions ($p > .05$). This suggests that when participants’ beliefs are weakened by corrective information, they simultaneously downgrade both their strength in these beliefs and their perceived relevance, without necessarily “propping up” alternative beliefs to fill this explanatory void.

Table 7: Effect of Interventions on Stated Relevance of Alternative Beliefs (Study 2)

	Mentioned Other Reason	Other Reason Relevance
Focal Belief Treatment	0.01 (0.02)	−0.05 (0.07)
Distal Belief Treatment	0.03 (0.02)	−0.02 (0.07)
Num.Obs.	1902	1902
R2	0.020	0.017
Covariates Included	✓	✓

+ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Argument recall Previous research on motivated reasoning has stressed the importance of biased recall, such that people more easily retrieve pro-attitudinal considerations than counter-attitudinal considerations (Taber and Lodge, 2006). While our design did not directly ask participants to provide pro-attitudinal arguments, we can test whether participants exhibit selective memory when recalling arguments that bear on their focal beliefs. We assess this in the second waves of Study 1 and 2 by asking participants to indicate whether they had seen two arguments drawn from their assigned condition and two from the other conditions each in the last month.⁸ Our evidence suggests slightly higher rates of recall for the focal versus distal belief condition in both studies, such that participants are 4pp more likely to recall an assigned argument in the focal condition ($\hat{ATE} = .09$; $SE = .02$; $p < .001$ in Study 1; $\hat{ATE} = .09$; $SE = .03$; $p < .001$ in Study 2) relative to the distal condition ($\hat{ATE} = .05$; $SE = .02$; $p < .05$; $\hat{ATE} = .05$; $SE = .02$ in Study 1; $p < .05$ in Study 2). This pattern of enhanced recall for focal belief counterarguments may help explain their greater persistence in producing attitude change relative to distal counterarguments. The heightened memorability of focal arguments suggests deeper engagement with information that challenges core justifications for attitudes, rather than a reflexive defensiveness against attitude-incongruent information.

Exploratory Analysis: The Factual Content in Citizens' Attitudes

Our findings suggest political attitudes respond consistently to newly learned information when the information addresses relevant beliefs. What if a person's most focal considerations are primarily evaluative in nature and absent factual content? For instance, a person who supports restrictive immigration policies due to negative perceptions of immigrants could treat either of the following as their most relevant consideration: a descriptive view that immigrants are responsible for violent crime, which can be directly addressed with available statistics; or an evaluation that immigrants are less deserving than citizens, which turns more

⁸For example, those assigned to the focal counterargument condition would see two statements from focal counterargument, two statements from the distal argument, and two placebo argument statements.

Table 8: Effect of Interventions on Argument Recall

	Study 1	Study 1	Study 1	Study 2	Study 2	Study 2
	Focal Recall	Distal Recall	Assigned Recall	Focal Recall	Distal Recall	Assigned Recall
Focal Correction	0.15*** (0.02)	0.04* (0.02)	0.09*** (0.02)	0.14*** (0.02)	0.06* (0.02)	0.09*** (0.03)
Distal Correction	0.003 (0.02)	0.11*** (0.02)	0.06* (0.02)	0.008 (0.02)	0.15*** (0.02)	0.05* (0.02)
Num.Obs.	2059	2063	2059	1660	1654	1655
R2	0.029	0.013	0.009	0.032	0.028	0.009
Covariates Included				✓	✓	✓

+ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

on preferences than on beliefs. Holding relevance constant, factual information targeting the latter consideration may produce less change in downstream immigration attitudes than information targeting the former.

While this is consistent with our framework, it raises the possibility of heterogeneous treatment effects across issues; respondents could be more likely to treat descriptive or factual justifications as focal for some issues than for others. A central thread in the public opinion literature is that Americans understand many issues primarily in symbolic or “easy” terms (Sears et al., 1980; Carmines and Stimson, 1980), meaning they have more to do with long-standing evaluative predispositions than factual beliefs about “hard” concrete policy benefits. Part of why numeracy information often fails to change immigration attitudes may be—as Hopkins, Sides and Citrin (2019) identify—that immigration is an issue for which purely symbolic preferences are especially focal (see also Kustov, Laaker and Reller, 2021).

Appendix B presents an exploratory analysis in which we categorize each respondent’s issue as “easy” or “hard” and their focal belief as primarily descriptive or evaluative. We find that although only around 44% of issues could be categorized as hard, nearly 70% of focal beliefs could be categorized as descriptive, with 63% of easy issue respondents and 79% of hard issue respondents providing focal beliefs that were primarily descriptive. Thus,

respondents frequently brought factual reasons to bear when justifying their deeply held attitude in conversation, even on “easy” issues. Our subsequent assessments of heterogeneous treatment effects offer suggestive evidence that informational arguments produced stronger attitude change for respondents who provided hard issues (descriptive beliefs) than for those with easy issues (evaluative beliefs). These results, though exploratory, help further reconcile the framework in this paper with canonical theories of public opinion, while underscoring the factual content in citizens’ attitudes and opportunity for facts to produce opinion change.

Discussion

Our findings shed new light on the relationship between beliefs and attitudes. By using LLMs to recover and confront beliefs that participants identified as central to their policy views, we provide experimental evidence that both “focal” and “distal” beliefs are capable of being corrected. Across two studies, participants updated their beliefs in the direction of the belief-incongruent information, running counter to expectations from approaches that strongly emphasize resistance or bias in the face of exposure to belief-incongruent information. The durability of these shifts—observed one week later—provides further support for a view of persuasion vis-à-vis beliefs that does not align neatly with approaches predicting extreme resistance to persuasion.

Focal counterarguments exerted larger and more durable effects on downstream attitudes than arguments targeting relatively peripheral beliefs. We observed stronger immediate reductions in attitudinal measures for participants who received a counterargument targeting beliefs they had explicitly described as key justifications for their position. In contrast, counterarguments to more distal beliefs that were unmentioned by the participant but nonetheless relevant to the attitude yielded smaller or inconsistent attitudinal effects. This dovetails with the notion that beliefs differ in their relevance to broader evaluations: changing a belief that participants see as important to their political views appears to produce

more discernible evidence of persuasion.

The greater memorability of focal counterarguments may help explain their persistence. Participants assigned to focal counterarguments recognized the arguments at higher rates, suggesting deeper engagement with these arguments. Throughout our studies, we found little evidence of widespread defensiveness in our studies. Participants in focal counterargument conditions did not uniformly strengthen alternative beliefs or discount the credibility of the argument. While we observed decreases in belief relevance, we did not find evidence that alternative considerations were “bolstered” to compensate for the weakened belief.

These findings help reconcile an apparent puzzle in the fact-checking literature: why does successful belief change often yield minimal attitude change? Our results suggest this pattern may stem partly from informational interventions that target malleable but relatively peripheral beliefs. Many popular fact-checks address viral claims on social media platforms but may not reflect the most relevant beliefs underlying individuals’ political attitudes. The methodological approach developed here—using LLMs to identify personally salient beliefs—provides a framework for designing more effective arguments by targeting beliefs that individuals themselves identify as central to their political evaluations.

Several questions merit further investigation. First, while we focused on strongly held political attitudes to provide a stringent test of our theoretical expectations, future work might examine whether the relative effectiveness of focal versus distal counterarguments generalizes to less crystallized attitudes. Second, our finding that arguments reduce both belief strength and perceived relevance raises questions about the relationship between these constructs and their distinct roles in attitude formation. Finally, additional research could explore whether providing multiple arguments targeting different types of beliefs produces cumulative effects on attitudes, as argued by Porter and Wood (2024*b*).

Our findings extend beyond fact-checking and informational interventions to piv-

otal questions about political cognition and attitude change, suggesting that citizens can meaningfully update both peripheral and central political beliefs when presented with belief-incongruent information. However, the translation of these changes in beliefs into attitude change appears to depend critically on the centrality of the corrected beliefs. This highlights the importance of understanding how individuals weight different beliefs in forming political attitudes and points toward strategies for more effective belief-based persuasion.

References

- Abascal, Maria, Tiffany J. Huang and Van C. Tran. 2021. "Intervening in Anti-Immigrant Sentiments: The Causal Effects of Factual Information on Attitudes toward Immigration." *The ANNALS of the American Academy of Political and Social Science* 697(1):174–191.
- Alt, James E., David D. Lassen and John Marshall. 2016. "Credible Sources and Sophisticated Voters: When Does New Information Induce Economic Voting?" *The Journal of Politics* 78(2):327–342.
- Arias, Eric, Horacio Larreguy, John Marshall and Pablo Querubín. 2022. "Priors Rule: When Do Malfeasance Revelations Help Or Hurt Incumbent Parties?" *Journal of the European Economic Association* 20(4):1433–1477.
- Bai, (Max) Hui, Jan G. Voelkel, Johannes C. Eichstaedt and Robb Willer. 2023. "Artificial Intelligence Can Persuade Humans on Political Issues."
- Barnes, Lucy, Avi Feller, Jake Haselswerdt and Ethan Porter. 2018. "Information, Knowledge, and Attitudes: An Evaluation of the Taxpayer Receipt." *The Journal of Politics* 80(2):701–706.
- Bartels, Larry M. 2002. "Beyond the Running Tally: Partisan Bias in Political Perceptions." *Political Behavior* 24(2):117–150.
- Brandt, Mark J. 2022. "Measuring the belief system of a person." *Journal of Personality and Social Psychology* 123(4):830.
- Brandt, Mark J. and G. Scott Morgan. 2022. "Between-person methods provide limited insight about within-person belief systems." *Journal of Personality and Social Psychology* 123(3):621–635.
- Bullock, John G. 2009. "Partisan Bias and the Bayesian Ideal in the Study of Public Opinion." *The Journal of Politics* 71(3):1109–1124.
- Carmines, Edward G. and James A. Stimson. 1980. "The Two Faces of Issue Voting." *American Political Science Review* 74(1):78–91.
- Chong, Dennis and James N. Druckman. 2007. "Framing Theory." *Annual Review of Political Science* 10(Volume 10, 2007):103–126.
- Citrin, Jack and John Sides. 2008. "Immigration and the Imagined Community in Europe and the United States." *Political Studies* 56(1):33–56.
- Clifford, Scott, Geoffrey Sheagley and Spencer Piston. 2021. "Increasing Precision without Altering Treatment Effects: Repeated Measures Designs in Survey Experiments." *American Political Science Review* 115(3):1048–1065.
- Converse, Philip E. 1964. The Nature of Belief Systems in Mass Publics. In *Ideology and Discontent*, ed. D. E. Apter. Ann Arbor: University of Michigan Press pp. 206–261.
- Cook, Fay Lomax, Lawrence R. Jacobs and Dukhong Kim. 2010. "Trusting What You

- Know: Information, Knowledge, and Confidence in Social Security.” *The Journal of Politics* 72(2):397–412.
- Coppock, Alexander. 2023. *Persuasion in Parallel: How Information Changes Minds about Politics*. Chicago Studies in American Politics Chicago, IL: University of Chicago Press.
- Coppock, Alexander, Kimberly Gross, Ethan Porter, Emily Thorson and Thomas J. Wood. 2023. “Conceptual Replication of Four Key Findings about Factual Corrections and Misinformation during the 2020 US Election: Evidence from Panel-Survey Experiments.” *British Journal of Political Science* 53(4):1328–1341.
- Dias, Nicholas C, Laurits F Aarslew, Kristian Vrede Skaaning Frederiksen, Yphtach Lelkes, Lea Pradella and Sean J Westwood. 2024. “Correcting misperceptions of partisan opponents is not effective at treating democratic ills.” *PNAS Nexus* 3(8):pgae304.
- Fearon, James D. 1999. Electoral Accountability and the Control of Politicians: Selecting Good Types versus Sanctioning Poor Performance. In *Democracy, Accountability, and Representation*, ed. Adam Przeworski, Bernard Manin and Susan C. Stokes. Cambridge Studies in the Theory of Democracy Cambridge: Cambridge University Press pp. 55–97.
- Festinger, Leon. 1957. *A Theory of Cognitive Dissonance*. Stanford University Press.
- Festinger, Leon, Henry W. Riecken and Stanley Schachter. 1956. *When Prophecy Fails*. Minneapolis, MN, US: University of Minnesota Press.
- Fishbein, M. 1979. “A theory of reasoned action: Some applications and implications.” *Nebraska Symposium on Motivation* 27:65–116.
- Fishbein, Martin and Icek Ajzen. 1975. *Belief, Attitude, Intention, and Behavior: An Introduction to Theory and Research*. Addison-Wesley Publishing Company.
- Fishbein, Martin and Icek Ajzen. 2010. *Predicting and Changing Behavior: The Reasoned Action Approach*. New York: Taylor & Francis.
- Fishman, Nic and Nicholas T. Davis. 2022. “Change We Can Believe In: Structural and Content Dynamics within Belief Networks.” *American Journal of Political Science* 66(3):648–663.
- Gerber, Alan and Donald Green. 1999. “Misperceptions about Perceptual Bias.” *Annual Review of Political Science* 2(Volume 2, 1999):189–210.
- Grigorieff, Alexis, Christopher Roth and Diego Ubfa. 2020. “Does Information Change Attitudes Toward Immigrants?” *Demography* 57(3):1117–1143.
- Helfers, Anna and Mirjam Ebersbach. 2023. “The differential effects of a governmental debunking campaign concerning COVID-19 vaccination misinformation.” *Journal of Communication in Healthcare* 16(1):113–121.
- Hill, Seth J. 2017. “Learning Together Slowly: Bayesian Learning about Political Facts.” *The Journal of Politics* 79(4):1403–1418.

- Hochschild, Jennifer L. and Katherine Levine Einstein. 2015. "Do Facts Matter? Information and Misinformation in American Politics." *Political Science Quarterly* 130(4):585–624.
- Hopkins, Daniel J., John Sides and Jack Citrin. 2019. "The Muted Consequences of Correct Information about Immigration." *The Journal of Politics* 81(1):315–320.
- Jerit, Jennifer. 2009. "How Predictive Appeals Affect Policy Opinions." *American Journal of Political Science* 53(2):411–426.
- Jervis, Robert. 2006. "Understanding Beliefs." *Political Psychology* 27(5):641–663.
- Key, V. O. 1949. *Southern Politics in State and Nation*. New York: A.A. Knopf.
- Kuklinski, James H., Paul J. Quirk, David W. Schwieder and Robert F. Rich. 1998. "'Just the Facts, Ma'am': Political Facts and Public Opinion." *The Annals of the American Academy of Political and Social Science* 560:143–154.
- Kuklinski, James H., Paul J. Quirk, Jennifer Jerit, David Schwieder and Robert F. Rich. 2000. "Misinformation and the Currency of Democratic Citizenship." *The Journal of Politics* 62(3):790–816.
- Kunda, Ziva. 1990. "The Case for Motivated Reasoning." *Psychological bulletin* 108(3):480.
- Kustov, Alexander, Dillon Laaker and Cassidy Reller. 2021. "The Stability of Immigration Attitudes: Evidence and Implications." *The Journal of Politics* 83(4):1478–1494.
- Kustov, Alexander and Michelangelo Landgrave. 2025. "Immigration is difficult?! Informing voters about immigration policy fosters pro-immigration views." *Journal of Experimental Political Science* pp. 1–13.
- Kuziemko, Ilyana, Michael I. Norton, Emmanuel Saez and Stefanie Stantcheva. 2015. "How Elastic Are Preferences for Redistribution? Evidence from Randomized Survey Experiments." *American Economic Review* 105(4):1478–1508.
- Lawrence, Eric D and John Sides. 2014. "The consequences of political innumeracy." *Research & Politics* 1(2):2053168014545414.
- Little, Andrew T. 2025. "How to Distinguish Motivated Reasoning from Bayesian Updating." *Political Behavior*.
- Lodge, Milton and Charles S. Taber. 2013. *The Rationalizing Voter*. Cambridge Studies in Public Opinion and Political Psychology Cambridge: Cambridge University Press.
- Mellon, Jonathan, Jack Bailey, Ralph Scott, James Breckwoldt and Marta Miori. 2022. "Does GPT-3 know what the Most Important Issue is? Using Large Language Models to Code Open-Text Social Survey Responses At Scale."
- Nelson, Thomas E. and Zoe M. Oxley. 1999. "Issue Framing Effects on Belief Importance and Opinion." *The Journal of Politics* 61(4):1040–1067.
- Nelson, Thomas E., Zoe M. Oxley and Rosalee A. Clawson. 1997. "Toward a Psychology of Framing Effects." *Political Behavior* 19(3):221–246.

- Nyhan, Brendan, Ethan Porter, Jason Reifler and Thomas J. Wood. 2020. "Taking Fact-Checks Literally But Not Seriously? The Effects of Journalistic Fact-Checking on Factual Beliefs and Candidate Favorability." *Political Behavior* 42(3):939–960.
- Oliver, J. Eric and Tali Mendelberg. 2000. "Reconsidering the Environmental Determinants of White Racial Attitudes." *American Journal of Political Science* 44(3):574–589.
- Petty, Richard E. 2024. "What Is Unique About Acceptance and Correction of Misinformation? Insights from Work on Attitudes, Persuasion, and Beyond." *American Psychologist* 79(9):1275–1288.
- Porter, Ethan and Thomas J Wood. 2024a. "Factual corrections: Concerns and current evidence." *Current Opinion in Psychology* 55:101715.
- Porter, Ethan and Thomas J. Wood. 2024b. "Factual corrections: Concerns and current evidence." *Current Opinion in Psychology* 55:101715.
- Porter, Ethan, Yamil R. Velez and Thomas J. Wood. 2022. "Factual Corrections Eliminate False Beliefs About COVID-19 Vaccines." *Public Opinion Quarterly* 86(3):762–773.
- Porter, Ethan, Yamil Velez and Thomas J Wood. 2023. "Correcting COVID-19 vaccine misinformation in 10 countries." *Royal Society open science* 10(3):221097.
- Sears, David O., Richard R. Lau, Tom R. Tyler and Harris M. Allen. 1980. "Self-Interest vs. Symbolic Politics in Policy Attitudes and Presidential Voting." *The American Political Science Review* 74(3):670–684.
- Slothuus, Rune. 2008. "More Than Weighting Cognitive Importance: A Dual-Process Model of Issue Framing Effects." *Political Psychology* 29(1):1–28.
- Stagnaro, Michael Nicholas, James Druckman, Adam J Berinsky, Antonio Alonso Arechar, Robb Willer and David Rand. 2024. "Representativeness versus attentiveness: a comparison across nine online survey samples." *PsyArXiv* 22.
- Taber, Charles S. and Milton Lodge. 2006. "Motivated Skepticism in the Evaluation of Political Beliefs." *American Journal of Political Science* 50(3):755–769.
- Velez, Yamil Ricardo and Patrick Liu. 2024. "Confronting Core Issues: A Critical Assessment of Attitude Polarization Using Tailored Experiments." *American Political Science Review* pp. 1–18.
- Walter, Nathan, Jonathan Cohen, R Lance Holbert and Yasmin Morag. 2020. "Fact-checking: A meta-analysis of what works and for whom." *Political communication* 37(3):350–375.
- Wong, Cara J. 2007. "'Little' and 'Big' Pictures in Our Heads: Race, Local Context, and Innumeracy about Racial Groups in the United States." *The Public Opinion Quarterly* 71(3):392–412.
- Zaller, John R. 1992. *The Nature and Origins of Mass Opinion*. Cambridge Studies in Public Opinion and Political Psychology Cambridge: Cambridge University Press.

Appendix

A Study Details	1
B Exploratory Analyses	17

A Study Details

A.1 Question Wording

Pre-treatment Measures

Issue Attitudes (7-point Likert scale) To what extent do you agree or disagree with the following statement? *[LLM-generated attitude statement based on conversation]*

1. Strongly disagree
2. Disagree
3. Somewhat disagree
4. Neither agree nor disagree
5. Somewhat agree
6. Agree
7. Strongly agree

Focal Belief Strength (4-point scale) How accurate do you think the following statement is? *[LLM-generated focal belief statement based on conversation]*

1. Not at all accurate
2. Slightly accurate
3. Moderately accurate
4. Very accurate

Distal Belief Strength (4-point scale) How accurate do you think the following statement is? *[LLM-generated distal belief statement based on conversation]*

1. Not at all accurate
2. Slightly accurate
3. Moderately accurate
4. Very accurate

Attitude Certainty (Study 2 only, 101-point scale) How certain are you regarding this position: *[LLM-generated attitude statement based on conversation]*

When giving a score, keep in mind:

- 100 = Complete certainty. You could not be more sure of where you stand.
- 75 $\approx$ Very sure about where you stand.
- 50 $\approx$ Pretty sure about where you stand.
- 25 $\approx$ Mildly sure about where you stand.

- 0 = Complete uncertainty. You have a belief, but you are very unsure if you should hold it.

Attitude Defense Scale (Study 2 only) To what extent do you agree or disagree with the following statements about your level of commitment to the following issue position:

[LLM-generated attitude statement based on conversation]

Please indicate your level of agreement ranging from ‘Strongly Disagree’ to ‘Strongly Agree.’

1. I would *not* be confident defending this position in a telephone interview that will end up in the local newspaper.
2. I could confidently defend this position in a filmed interview for live national television.
3. I would *not* be confident defending this position in a press conference in front of news reporters.
4. I could confidently defend this position in a public speech on a college campus.
5. I would *not* be confident defending this position at a city council meeting.

Response options:

1. Strongly disagree
2. Disagree
3. Somewhat disagree
4. Neither agree nor disagree
5. Somewhat agree
6. Agree
7. Strongly agree

Post-treatment Measures

Attitude Extremity To what extent do you agree or disagree with the following statements about your level of commitment to the following issue position:

[LLM-generated attitude statement based on conversation]

Please indicate your level of agreement ranging from ‘Strongly Disagree’ to ‘Strongly Agree.’

- I would phone state representatives to advocate for this issue.
- I would risk personal harm or even death to advocate for this cause.
- I would initiate grassroots campaigns to mobilize supporters for this issue.
- I would walk five miles to bring attention to this issue.
- I would join a protest, despite a risk of arrest, to support this issue.

- I *wouldn't* cross the street to participate in a rally concerning this issue.
- I would go on a hunger strike to amplify awareness of this issue.
- I would willingly get fired from my job to advance this issue.

Belief Relevance (4-point scale) Consider the reasons why you hold this view:
[LLM-generated attitude statement based on conversation]

How important are the following ideas for why you believe in this view?
[LLM-generated focal and distal belief statements]

1. Not at all important
2. Low importance
3. Moderately important
4. Very important

Alternative Beliefs (Study 2 only, text entry) If there are **more important reasons** why you hold this view that are not listed, **please tell us using the “Other” text box.**
[Open text response]

Wave 2 Measures (Approximately One Week Later)

Argument Recall Test In the last month, have you seen any of the following arguments? Please check all that apply.
[Randomly selected sentences from focal counterargument, distal counterargument, and placebo conditions]

1. Yes, I have seen this argument
2. No, I have not seen this argument

Attitude and Belief Measures All attitude strength, attitude extremity, belief strength, belief relevance, and attitude defense measures from Wave 1 were repeated in Wave 2.

A.2 Example Inputs and Completions

Example 1: Providing universal health care

Issue Position Providing universal health care

Chat History (chatHistory)

- **Assistant:** Thank you for sharing your stance on providing universal health care. That's an important and impactful issue. Could you elaborate on why you believe universal health care is necessary? What are the key reasons or beliefs that support your position on this topic?
- **User:** I think the right to health and physical safety is innate to humankind
- **Assistant:** That's a compelling belief. You mentioned that the right to health and physical safety is innate to humankind. Could you explain a bit more about how you see this right being realized through universal health care? What specific aspects of universal health care do you think are essential to uphold this right?
- **User:** Maslow's hierarchy of need focuses around people being housed and not physically being in danger. Universal healthcare helps solve an important part of that
- **Assistant:** That's a thoughtful connection to Maslow's hierarchy of needs. It seems you see universal health care as a foundational element for individuals to feel secure and safe, which is crucial for their overall well-being. How do you think access to universal health care impacts other levels of needs in Maslow's hierarchy, such as psychological or social needs? Additionally, are there specific outcomes you believe would result from implementing universal health care that could further address these needs?
- **User:** Being healthy or not being in fear of going to seek a health practitioner can really help mental health
- **Assistant:** Absolutely, the connection between physical health and mental health is significant. It sounds like you believe that universal health care could reduce the anxiety and fear associated with accessing medical services, which in turn could support better mental health outcomes. Can you elaborate on how you think universal health care could contribute to improved mental health for individuals? Are there specific examples or scenarios you envision where this would be particularly impactful?

Structured Data

Elicited Attitude: Universal health care is necessary.

Focal Belief: Universal health care addresses basic human needs according to Maslow's hierarchy.

Distal Belief: Universal health care reduces health disparities among different socioeconomic groups.

Focal Counterargument: The belief that universal health care directly addresses basic human needs as outlined in Maslow's hierarchy of needs is complicated by the multifaceted nature of well-being and safety. While the hierarchy does prioritize safety and health, the implementation of universal health care does not automatically guarantee these outcomes. Various countries with universal health care systems still face significant challenges in ensuring timely and quality access to services. For instance, lengthy wait times for medical procedures are a common issue, as seen in countries like Canada and the UK, where patients sometimes wait months for critical surgeries. These delays can exacerbate health issues and create additional stress, potentially undermining the very safety and security that universal health care aims to provide.

Furthermore, universal health care systems often face financial sustainability challenges, which can lead to rationing of care or reduction in available services. The economic strain of maintaining such systems can result in increased taxation or reallocation of funds from other critical social services, inadvertently affecting the broader social safety net. This can undermine the holistic approach to meeting human needs, as resources may be diverted away from education, housing, or employment programs, which are also crucial for security and self-actualization.

Additionally, the assumption that universal health care alone can fulfill the need for health and safety overlooks the importance of social determinants of health, such as nutrition, housing, and environmental factors. Addressing these determinants requires a multifaceted approach that extends beyond the scope of health care systems. Comprehensive public health strategies and policies that integrate health care with broader social initiatives are necessary to create an environment where all individuals can truly achieve safety and security. Without these additional supports, universal health care may not be sufficient to meet the foundational human needs described by Maslow.

Distal Counterargument: While the aim of universal health care is to reduce health disparities, evidence suggests that merely providing access does not automatically equalize health outcomes across different socioeconomic groups. Research indicates that disparities persist even in countries with established universal health care systems. For example, despite having access to health care, individuals from lower socioeconomic backgrounds often experience worse health outcomes compared to their wealthier counterparts, due to factors such as health literacy, lifestyle, and access to preventative care.

Socioeconomic disparities in health are deeply rooted in structural inequalities that universal health care alone cannot address. Factors such as education, income, employment, and environment play a critical role in determining health outcomes. A study published in *The Lancet* highlighted that even in countries with universal health care, mortality rates are higher among lower-income groups due to these social determinants. Addressing these broader social issues requires targeted policies beyond health care, such as improving education and employment opportunities, which can help bridge the gap in health outcomes.

Moreover, cultural and geographic barriers can still limit access to health care services for marginalized communities within universal systems. In rural or remote areas, shortages of medical professionals and facilities can restrict access to timely and adequate care, perpetuating health disparities. Policy interventions need to focus on improving infrastructure and incentivizing health professionals to work in underserved areas to ensure equitable access for all.

Placebo Argument: The biannual adjustment of daylight savings time (DST) is a subject of debate, with some suggesting it disrupts sleep patterns and health. However, studies have shown that the impact of DST on sleep is minimal and transient. Research published in the journal *Sleep Medicine Reviews* indicates that while there may be a temporary disruption in circadian rhythms, most individuals adapt within a few days. This adaptation is facilitated by the body's natural ability to adjust to changes in light exposure, which is a stronger determinant of circadian rhythms than the one-hour shift of the clock.

Furthermore, DST has been associated with positive societal impacts, such as energy savings and reduced crime rates. The U.S. Department of Energy reported that DST leads to a reduction in energy consumption by approximately 0.03% annually, as extended daylight hours reduce the need for artificial lighting and heating in the evenings. Additionally, a study in the *Review of Economics and Statistics* found that the additional hour of daylight in the evening correlates with a reduction in robbery rates, as increased natural light can deter criminal activity.

While there are concerns regarding potential health impacts, the benefits of DST in terms of energy efficiency and public safety present a compelling case for its continuation. Policymakers must weigh these advantages against the minimal disruptions to circadian rhythms, considering that the overall societal benefits may outweigh the short-lived personal inconveniences.

Example 2: Prioritizing policies that reduce inflation

Issue Position Prioritizing policies that reduce inflation

Chat History (chatHistory)

- **Assistant:** Thank you for sharing your perspective on prioritizing policies that reduce inflation. That's a significant issue that affects many aspects of our economy and daily lives. To better understand your stance, could you elaborate on why you believe reducing inflation is so important? What specific beliefs or values guide your position on this issue?
- **User:** The high cost of goods and high cost of borrowing is crippling many Americans.
- **Assistant:** That's a compelling point. It sounds like you believe that high inflation

directly impacts the financial well-being of individuals and families. Could you share more about how you think reducing inflation would improve the situation for Americans? Are there particular aspects of daily life or the economy that you think would benefit most from lower inflation?

- **User:** The lower the cost of the goods the more money Americans to spend which is good for our economy
- **Assistant:** Absolutely, it seems you see a direct connection between lower costs and increased consumer spending, which can drive economic growth. Can you elaborate on how you believe this cycle works? What specific outcomes do you think would arise from reduced inflation in terms of consumer behavior and overall economic health?
- **User:** Reduced cost of goods will provide more opportunities for growth in our economy. Meaning more jobs, stronger dollar, and financial ease for many.
- **Assistant:** That makes a lot of sense. It sounds like you view reduced inflation not just as a relief for consumers, but as a catalyst for broader economic growth and stability. Could you share more about how you think these changes—like job growth and a stronger dollar—would come about? Are there specific policies or measures you believe should be prioritized to achieve these outcomes?

Structured Data

Elicited Attitude: Reducing inflation is important for economic growth.

Focal Belief: Reduced inflation leads to more consumer spending, thus boosting the economy.

Distal Belief: A stronger dollar resulting from reduced inflation benefits the U.S. economy by increasing purchasing power abroad.

Focal Counterargument: While the notion that lower inflation should lead to increased consumer spending might seem intuitive, the reality is more complex. Firstly, consumer behavior does not solely hinge on inflation rates. Historical data indicates that consumer confidence and spending are often influenced by a variety of factors, such as employment security and wage growth. For instance, during the Great Recession, despite low inflation, consumer spending was significantly depressed due to high unemployment and stagnant wages. Additionally, the Phillips Curve, an economic concept that historically suggested a trade-off between inflation and unemployment, has in recent times shown signs of breakdown. This indicates that reducing inflation does not automatically equate to more consumer spending or economic growth. Moreover, low inflation can sometimes reflect weak demand, which could signal economic stagnation rather than growth.

Furthermore, empirical studies have shown that in certain cases, deflation, or negative inflation, can be detrimental to the economy. The Japanese economy during the 1990s, known as the Lost Decade, suffered from deflationary pressures that led to stag-

nation despite efforts to stimulate spending. The psychological impact of deflation can cause consumers to delay purchases, expecting prices to fall further, which in turn can suppress economic activity. This highlights the complexity of the relationship between inflation rates and economic health, suggesting that simply reducing inflation is not a panacea for economic woes.

Additionally, while reducing inflation may alleviate the immediate cost pressures on households, it does not address structural issues such as income inequality and wage stagnation. These factors have a profound impact on consumer spending power. Reports from the Economic Policy Institute and other research bodies have consistently shown that income growth for the top earners significantly outpaces that of the average worker, which means that even with lower inflation, many Americans may not feel the benefit in their purchasing power. Therefore, policy approaches need to be multifaceted, addressing both inflation and these underlying structural issues to truly enhance consumer spending and economic growth.

Distal Counterargument: The belief that a stronger dollar inherently benefits the U.S. economy by increasing purchasing power abroad is only partially accurate and overlooks several key economic dynamics. While a stronger dollar can indeed make foreign goods and services cheaper for U.S. consumers, it also makes American exports more expensive for foreign buyers. This can negatively impact U.S. manufacturers and exporters, potentially leading to trade deficits and job losses in export-driven industries. Historical data from the U.S. Bureau of Economic Analysis shows that periods of a robust dollar often correlate with increased trade deficits, as seen in the early 1980s and mid-2010s. This undermines the competitiveness of American products in the global market, which can have ripple effects throughout the economy.

Moreover, a stronger dollar can lead to capital inflows into the U.S. as foreign investors seek to capitalize on higher returns. While this might seem beneficial at first glance, it can lead to asset bubbles, particularly in real estate and equities, as was observed in the lead-up to the 2008 financial crisis. When these bubbles burst, they can cause significant economic disruptions, negating any short-term benefits of increased purchasing power. The International Monetary Fund has cautioned that countries with rapidly appreciating currencies can face challenges related to financial stability, highlighting the nuanced effects a stronger dollar can have on the broader economy.

Additionally, the impact of a strong dollar on the global economy can have feedback effects on the U.S. economy itself. Emerging markets, which often have debt denominated in dollars, can struggle under the weight of a stronger dollar, leading to economic instability in these regions. Such instability can affect global supply chains and demand for U.S. goods, ultimately impacting U.S. economic growth. The interconnected nature of the global economy means that the benefits of a stronger dollar are not straightforward and can lead to unintended negative consequences both domestically and internationally.

Placebo Argument: Adjustments for daylight savings time have long been a subject of de-

bate, often criticized for disrupting schedules and reducing productivity. However, a wealth of research suggests that the benefits, particularly in terms of energy conservation and public health, are substantial. According to a study by the U.S. Department of Energy, daylight savings time results in a significant reduction in energy consumption, amounting to about 0.5% per day. While this may seem modest, when aggregated across the entire U.S. population, it translates to a considerable amount of energy savings, contributing to decreased reliance on non-renewable energy sources and a reduction in carbon emissions.

Moreover, daylight savings time has been associated with increased outdoor activity, as extended daylight hours encourage people to engage in physical activities after work. This can lead to improvements in public health by reducing the prevalence of lifestyle-related diseases such as obesity and cardiovascular issues. A study published in the *Journal of Environmental Psychology* found that the extra hour of daylight in the evening is linked to an increase in outdoor recreational activities, which promotes better mental and physical health outcomes.

Furthermore, daylight savings time can have economic benefits by boosting retail sales. Longer daylight in the evening encourages consumers to shop and dine out, supporting local businesses and contributing to economic growth. This is particularly relevant for the tourism and hospitality industries, which benefit from extended daylight hours. A report by the National Association of Convenience Stores highlighted that daylight savings time contributes to billions of dollars in increased sales annually. Thus, while the transition can be inconvenient, the broader benefits of daylight savings time adjustments are significant and multifaceted.

Example 3: Addressing public infrastructure

Issue Position For me, one of the issues that I pay particularly close attention to is public infrastructure. This includes the ongoing maintenance of the roads, bridges, signage, and train systems that I use. It is important that constant maintenance is performed on these systems to ensure that they are safe and up to modern standards. Ensuring that they are constantly maintained will make it easier for everybody to travel to where they need to be.

Chat History (chatHistory)

- **Assistant:** Thank you for sharing your thoughts on public infrastructure. It sounds like you see it as a critical issue. Can you tell me more about why you consider the maintenance of roads, bridges, signage, and train systems to be so important? What specific beliefs or experiences have shaped your stance on this issue?
- **User:** I want to make sure that the roads, signage and train system in my city function as they should. My beliefs are shaped by my day to day experience with these systems.
- **Assistant:** It's great to hear that your beliefs are grounded in your everyday experiences. Can you elaborate on any specific experiences that have influenced your views on the importance of maintaining these public infrastructure systems? Were there any particular incidents or observations that made you realize how crucial this issue is?
- **User:** Yes, for example, I have encountered potholes on the road and train delays that have impacted my commute
- **Assistant:** Those experiences certainly sound frustrating and can highlight the importance of well-maintained infrastructure. How do you think issues like potholes and train delays affect not just your personal commute but also the wider community? What consequences do you see arising from neglecting infrastructure maintenance?
- **User:** When issues like these are not maintained, they increase traffic, delays, and cause great disruption to every day people's lives.
- **Assistant:** That's a compelling point. It seems like you believe that inadequate maintenance can lead to broader societal issues. How do you think these disruptions impact the economy or the quality of life for residents in your community? Are there any specific economic or social outcomes that concern you?

Structured Data

Elicited Attitude: Proper maintenance of public infrastructure is essential.

Focal Belief: Inadequate maintenance of roads and train systems causes traffic and delays.

Distal Belief: Poor infrastructure leads to economic downturns by hampering trade and productivity.

Focal Counterargument: While it is intuitive to associate poor road and train maintenance with increased traffic and delays, the relationship is more nuanced. Comprehensive studies, such as those conducted by the U.S. Department of Transportation, suggest that traffic congestion often results from a variety of factors, including urban planning, population density, and public transportation availability, rather than solely infrastructure maintenance. For instance, cities with robust public transportation systems but infrequent maintenance still manage to alleviate congestion due to effective urban planning and diversified transport options. Furthermore, the perception of increased delays may stem more from increased vehicle usage and urban sprawl, which outpaces infrastructure capacity, than from maintenance lapses alone.

Moreover, the role of technology and smart traffic management systems has been shown to mitigate congestion even in areas where infrastructure maintenance is less than ideal. Research from the European Commission on Transport reveals that implementing adaptive traffic signals and real-time traffic monitoring can significantly reduce congestion without immediate maintenance improvements. Therefore, while maintenance plays a role, addressing traffic and delays effectively demands a multi-pronged approach that includes technological innovations and strategic planning, not just maintenance.

Additionally, the economic impact of infrastructure maintenance is complex. While immediate maintenance might seem costly, strategic investments in infrastructure that incorporate modern materials and smart technologies can yield long-term savings and efficiency gains. Studies from the World Bank highlight that infrastructure investments that focus on resilience and adaptability tend to perform better economically over time, suggesting that the focus should perhaps be broader than maintenance alone.

Distal Counterargument: The assumption that poor infrastructure directly leads to economic downturns is reductive, as the relationship between infrastructure and economic performance is mediated by numerous other factors. According to a World Economic Forum report, while infrastructure is a significant component of a nation's economic framework, other elements such as policy stability, workforce education, and technological adoption often play more decisive roles in determining economic outcomes. For instance, countries like Singapore have excelled economically with targeted infrastructure investments that coincide with robust regulatory frameworks and skilled labor forces, suggesting a model where infrastructure supports rather than drives economic success.

Historical analysis further complicates the direct causality between infrastructure quality and economic performance. During the post-World War II era, countries like Germany and Japan experienced rapid economic growth despite initially poor infrastructure conditions. This growth was primarily attributed to policy reforms, innovation, and investment in human capital rather than immediate infrastructure improvements. Thus, while infrastructure contributes to economic activity, it is not the sole or even primary determinant of economic health.

Moreover, the introduction of digital infrastructure and technology-driven economies has reshaped the traditional understanding of infrastructure's role. According to research by the MIT Digital Economy Lab, digital infrastructure can compensate for physical infrastructure deficiencies by enabling remote work, enhancing logistic efficiencies, and fostering global connectivity. This indicates that modern economic resilience might rely more on digital advancements than on traditional infrastructure maintenance.

Placebo Argument: The belief that daylight savings time adjustments are detrimental due to their disruptive effects overlooks several positive outcomes associated with the practice. Firstly, daylight savings time (DST) was initially implemented to reduce energy consumption by taking advantage of natural daylight during evening hours. Studies conducted by the U.S. Department of Energy have shown that DST results in a modest reduction in energy usage, as people use less electricity for lighting and appliances during the longer daylight hours.

Additionally, DST has been linked to a decrease in road accidents. Research published in the American Economic Journal found that the shift to DST correlates with a reduction in traffic fatalities due to increased daylight during peak traffic hours. This safety benefit arises from improved visibility and a reduction in the likelihood of accidents caused by fatigue during evening commutes.

Furthermore, DST can have positive effects on public health. An analysis by the Journal of Environmental Psychology suggests that the additional daylight encourages outdoor activities and physical exercise, contributing to better overall health. While the transition into and out of DST can cause temporary disruptions to sleep patterns, these are generally short-lived and do not outweigh the broader societal benefits. Therefore, while there are challenges associated with DST, its advantages in energy conservation, public safety, and health promotion provide compelling reasons for its continued implementation.

Example 4: Reducing immigration

Issue Position cost of living and the immigration at the border. also inflation

Chat History (chatHistory)

- **Assistant:** Thank you for sharing that you care about the cost of living, immigration at the border, and inflation. Let's dive deeper into these issues. First, could you tell me which of these issues you consider the most important and why? What specific aspects of this issue resonate with you the most? Additionally, what beliefs or experiences have shaped your attitude toward this issue? For example, do you have particular concerns about how it affects your community or your personal life?

- **User:** immigration because it is watering down the very fabric of our society where as all people are welcomed but in a structured way with the acclamations' in tact.
- **Assistant:** Thank you for sharing your perspective on immigration. It sounds like you believe that immigration should be managed in a way that preserves the foundational aspects of society. Can you elaborate on what you mean by “watering down the very fabric of our society”? What specific elements of society do you feel are at risk, and how do you think structured immigration could address those concerns? Additionally, are there any particular policies or changes that you believe would help create a more structured immigration system?
- **User:** diluting the core beliefs that this country was founded upon. The constitution is a special document that needs to be adhered to as much as possible
- **Assistant:** I appreciate your thoughts on the importance of the Constitution and the foundational beliefs of the country. It seems that you believe that immigration can challenge those core beliefs. Could you expand on what specific core beliefs or values you feel are being diluted? How do you think a structured approach to immigration could help maintain those beliefs? Also, are there particular constitutional principles or rights that you think are most at risk due to current immigration practices?
- **User:** Basic language which is english, and the lack of the assimilation of the immigrants to converse and learn the specific ways of our society. Trying to make the population as homogenized as possible
- **Assistant:** Thank you for clarifying your stance. It sounds like you place a strong emphasis on the importance of a common language—specifically English—and the need for immigrants to assimilate into the societal norms and culture. Could you elaborate on why you believe that language and cultural assimilation are crucial for maintaining the integrity of society? How do you think these factors contribute to social cohesion and a shared identity? Furthermore, what specific measures do you think could encourage or facilitate language learning and cultural assimilation for immigrants?

Structured Data

Elicited Attitude: Immigration should be structured.

Focal Belief: Immigration dilutes core beliefs and language.

Distal Belief: Immigration increases economic inequality.

Focal Counterargument: The belief that immigration dilutes core beliefs and language overlooks the dynamic nature of cultural evolution and the historical context of immigration in the United States. Historically, waves of immigrants have contributed to the cultural, economic, and social fabric of the nation. For instance, during the early 20th century, large numbers of Italian, Irish, and Eastern European immigrants arrived in the U.S., each bringing distinct languages and customs. Over time, these groups integrated into American society, enriching it with cultural diversity while also adopting English and contributing to the American way of life. This process of cultural

exchange and adaptation is a hallmark of American identity, demonstrating that cultural integration and diverse backgrounds can coexist and strengthen societal cohesion.

Furthermore, studies have shown that immigrants are keen to learn English as it opens up opportunities for better employment, education, and social integration. According to the Migration Policy Institute, nearly half of the immigrants entering the U.S. are proficient in English, and language acquisition programs have shown significant success in improving language skills among non-native speakers. Government and community initiatives to provide accessible language learning resources further facilitate this process, ensuring that immigrants can effectively participate in society while maintaining their cultural heritage.

The notion of cultural dilution also underestimates the ability of societies to adapt and incorporate new influences. Cultural evolution is a continuous process, where new ideas and practices enrich the existing cultural landscape. The United States, with its history of diverse immigrant contributions, exemplifies how cultural diversity fosters creativity, innovation, and resilience. Rather than viewing cultural influences as threats, they can be seen as opportunities for growth and enrichment, enhancing the nation's global competitiveness and cultural richness.

Distal Counterargument: The belief that immigration increases economic inequality fails to consider the multifaceted economic contributions that immigrants provide to the host country. Immigrants often fill essential roles in the labor market, especially in industries facing labor shortages, such as agriculture, healthcare, and technology. According to the National Academy of Sciences, immigrants contribute positively to the economy by boosting productivity, driving innovation, and complementing the native workforce. They start businesses at a higher rate than native-born individuals, creating jobs and stimulating economic activity. Therefore, rather than exacerbating inequality, immigration can help alleviate economic disparities by fostering economic growth and expanding the labor market.

Moreover, the wages of native workers are generally not negatively affected by immigration. Research indicates that while there might be short-term wage adjustments in specific sectors, the long-term effect of immigration on native wages is negligible. In some cases, immigration has been shown to increase wages for native workers by allowing them to specialize in higher-skilled jobs while immigrants take on lower-skilled positions. This complementary dynamic enhances overall productivity and economic output, contributing to a more prosperous economy.

Additionally, the notion that immigration increases economic inequality overlooks the significant fiscal contributions immigrants make to the economy. Immigrants pay taxes, contribute to social security, and support local economies through consumption. The net fiscal impact of immigrants is generally positive, especially when considering the contributions of second-generation immigrants, who tend to achieve higher educational and income levels. By fostering economic dynamism and addressing demographic chal-

lenges such as an aging population, immigration can be a powerful tool for reducing economic inequality and ensuring long-term economic sustainability.

Placebo Argument: The belief that daylight savings time adjustments are beneficial overlooks significant evidence suggesting they may cause more harm than good. Originally implemented to conserve energy during World War I and II, daylight savings time (DST) was based on the assumption that aligning waking hours with daylight would reduce energy consumption. However, modern research has shown that the energy savings are minimal. A study by the U.S. Department of Energy found that DST reduces total energy use by only about 0.03%, a negligible amount given the complexities of modern energy consumption patterns.

Furthermore, the impact of DST on health and well-being has raised concerns. The abrupt change in time can disrupt human circadian rhythms, leading to sleep deprivation and associated health risks. Studies have linked DST transitions to an increase in heart attacks, workplace accidents, and road incidents in the days following the clock change. This disruption can have significant public health implications, outweighing any minor energy savings achieved through the policy.

Economically, the benefits of daylight savings time are also questionable. While some industries, such as retail and tourism, might experience slight increases in consumer activity due to extended daylight hours, others, like agriculture, face challenges. The shift in time can disrupt farming schedules and livestock routines, leading to inefficiencies and potential economic losses. As such, the supposed economic advantages of DST do not hold uniformly across sectors, suggesting that the practice may no longer be justified in the contemporary context.

A.3 Sample Composition

Below are sample composition details using demographic variables provided by CloudResearch Connect (Study 1) and Prolific (Study 2). For study 1, household income (HHI) is an ordinal variable with the following 17 factor levels: “Less than \$10,000,” “\$10,000–\$19,999,” “\$20,000–\$29,999,” ..., “\$100,000–\$124,999,” “\$125,000–\$149,999,” “\$150,000–\$174,999,” “\$175,000–\$199,999,” “\$200,000–\$249,999,” and “\$250,000 or more.” Education has 8 levels ranging from “Less than a high school diploma” to “Doctorate degree.”

Table A1: Means Estimates for Demographic Categories.

	Study 1 (CloudResearch)	Study 2 (Prolific)
Female	0.51	0.51
Age	42.92	46.13
HHI	7.95	Not provided
Education	4.44	Not provided
Democrat	0.42	0.29
Republican	0.22	0.28
Independent	0.29	0.43
Hispanic	0.11	Not provided
White	0.70	0.71
Black	0.12	0.15
AAPI	0.05	0.04
Mixed	Not provided	0.06

B Exploratory Analyses

B.1 Why Does Identifying Focal Beliefs Matter?

To examine how numeracy stacks up against other immigration-related beliefs with regard to relevance, we identified 167 respondents in study 1 and 139 respondents in study 2 whose strongly-held attitude expressed a clear position in favor of restricting immigration to the U.S. We then hand-coded the focal beliefs summarized from respondent’s open-ended discussions.⁹ As Table A2 illustrates, only 1.8% of study 1 respondents and 3.6% of study 2 respondents personally mentioned the size of the immigrant population as a reason for their attitudes.¹⁰ By contrast, the vast majority of respondents cited beliefs that unauthorized immigrants were responsible for increasing crime, taking public resources (e.g., healthcare, social services, housing) from citizens, or hurting the economy and taking jobs—precisely the misperceptions for which corrections have been shown to influence downstream immigration attitudes (Abascal, Huang and Tran, 2021).

Table A2: Focal Beliefs for Respondents in Favor of Restricting Immigration.

Focal Belief	Study 1 ($N = 167$)	Study 2 ($N = 139$)
Safety/Crime	35.3%	50.4%
Public resources	35.3%	23.7%
Economy/Jobs	22.2%	10.8%
Drugs	8.4%	4.3%
Rule of law	6.6%	12.2%
Disrupting society/Not assimilating	3.6%	1.4%
Unfair to legal immigrants	2.4%	1.4%
Numeracy (too many immigrants)	1.8%	3.6%
Voting illegally for Democrats	1.2%	1.4%
Some asylum seekers are not real refugees	0.6%	0.0%
U.S. is not responsible for issues abroad	0.6%	0.0%
Disease	0.6%	0.0%

One critique of our conversation-based approach to identifying relevant beliefs may be that respondents’ justifications are disingenuous, molded by social desirability. That respondents have organically offered their immigration attitude as a deeply held preference—no less, as an open-ended response in a majority of cases—should mitigate such concerns. A second critique may be that people are ill-aware of the true reasons for their attitudes, and therefore cannot identify their focal beliefs even if they mean to. If true, then empirically we should find no difference in the attitudinal effects of focal belief counterarguments and distal belief counterarguments. We test this question in our main results and find that focal belief counterarguments more strongly affect attitudes than distal belief counterarguments.

⁹For both studies, we filtered respondents’ elicited attitudes for the “immigra” stem. Of the remaining observations, 35 were pro-immigration and 3 were removed because they did not express a clear position in favor of increasing or limiting immigration. When hand-coding focal beliefs, some observations were coded as expressing more than one belief.

¹⁰In study 1, there were also three respondents whose deeply held attitude made explicit reference to numeracy, e.g., “There are too many immigrants.” There were two such respondents in study 2.

B.2 How Issues Vary in Descriptive Versus Evaluative Content

Compatible with our framework is the possibility that the considerations a person treats as most relevant to their attitude inherently possess little factual content, limiting the efficacy of factual counterarguments. By extension, respondents may be more likely to treat descriptive or factual justifications as focal for some issues than for others. Carmines and Stimson (1980) famously argued that issue voting can take one of two forms. Where so-called *hard*-issue voting involves a careful calculation of policy benefits, *easy*-issue voting is characterized by gut responses to political symbols. Though the same issue can be categorized as easy or hard depending on how it is framed and deliberated, easy issues typically are “symbolic rather than technical,” “more likely deal with policy ends than means,” and have been “long on the political agenda” (80).

We can assess whether the nature of respondents’ issues and beliefs moderate treatment effects by categorizing the issues elicited from respondents as easy or hard and categorizing beliefs as descriptive or evaluative. If, within our design, respondents provide primarily evaluative judgments rather than factual arguments when asked by the chatbot to justify their attitude, then this is likely to be reflected within the focal belief statements synthesized by GPT. For each respondent across studies 1 and 2, we used GPT-4.1 nano to code whether the focal and distal beliefs were primarily descriptive or evaluative and whether the deeply held attitude concerned an easy or hard issue. We used few-shot learning, providing several example statement-code pairs in the prompt. We then randomly sampled 50 respondents from each study and hand-coded beliefs and attitudes.

Table A3: Content of Focal and Distal Beliefs about Easy versus Hard Issues

	Focal Belief is Descriptive	Distal Belief is Descriptive
Intercept	0.63*** (0.01)	0.64*** (0.01)
Hard Issue	0.16*** (0.01)	0.13*** (0.01)
Num.Obs.	4418	4418
R2	0.030	0.019

+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

For the easy/hard issue categorization, the LLM coding matched the hand-coding in 92 of the 100 cases. With the descriptive/evaluative belief categorization, we found a 96% match for focal beliefs and 88% match for distal beliefs.

69.9% of focal beliefs were coded as descriptive by the LLM, and 43.7% of issues were coded as hard. Table A3 displays the results of a simple linear model regressing whether the focal or distal belief was categorized as descriptive (binary) on whether the elicited issue was categorized as hard (binary). Hard issues seem substantially more likely to be associated with descriptive beliefs. Focal beliefs were coded as descriptive for 79% of respondents whose elicited issue was hard, compared to 63% for respondents whose elicited issue was easy.

To examine whether the efficacy of informational arguments varies with the easy or hard nature of the respondent's issue, we assess heterogeneous treatment effects (HTEs). We regress attitude strength outcomes on pre-treatment attitude strength and the interaction of treatment condition with issue/belief categorization. Figure A1 displays HTEs by issue using estimated marginal means, and Table A4 shows model results with the interactions. In both studies, focal counterargument exposure decreased attitude strength relative to the placebo condition for respondents with issues categorized as easy. However, both focal and distal counterarguments were more effective for respondents with issues categorized as hard in studies 1 and 2.

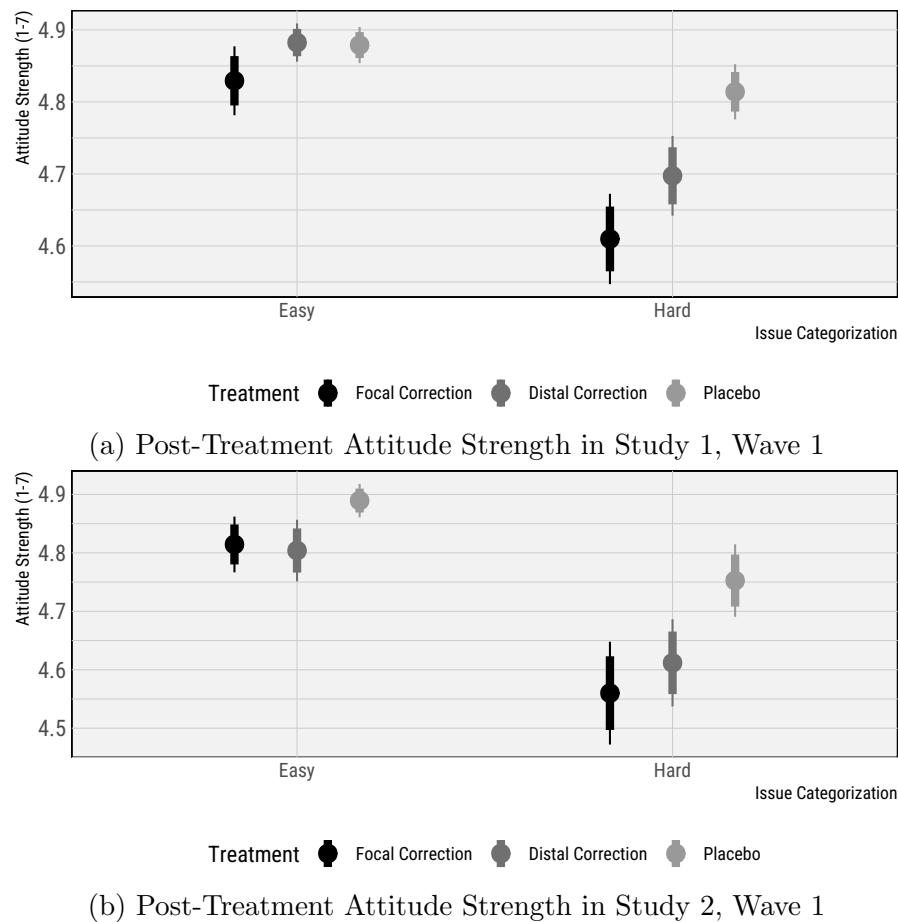
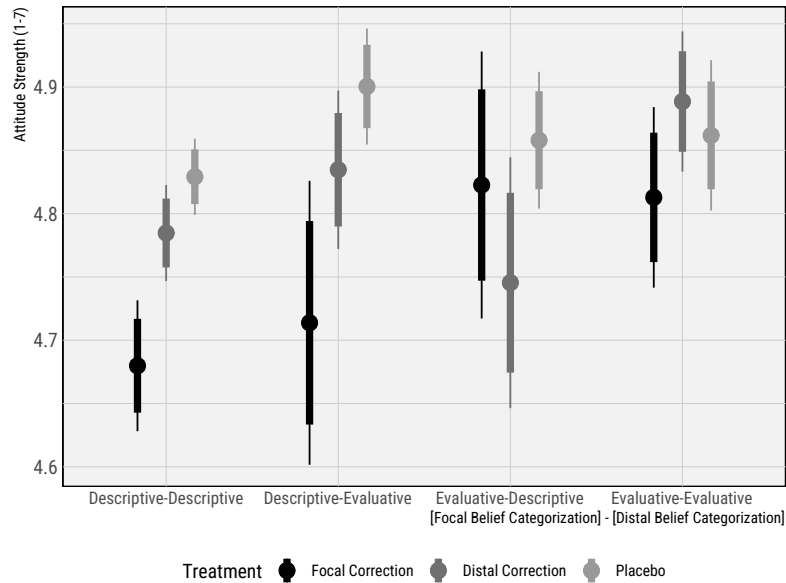
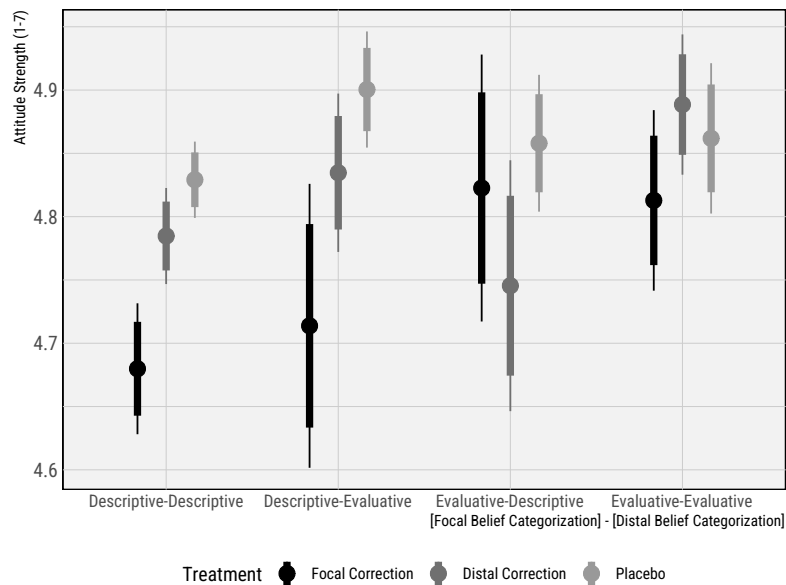


Figure A1: HTEs by Issue Categorization (Easy/Hard)

Figure A2 examines treatment effect heterogeneity by belief categorization. We divided respondents into four subgroups; for instance, “Descriptive-Evaluative” refers to the respondent's focal belief being categorized as primarily descriptive and distal belief being categorized as primarily evaluative. For study 1 (panel a), comparing columns 1-2 to columns 3-4 suggests the focal treatment was more effective when the focal belief was categorized as primarily descriptive as opposed to primarily evaluative. The difference is less clear for study 2 (panel b). Table A5 displays the interaction effects.



(a) Post-Treatment Attitude Strength in Study 1, Wave 1



(b) Post-Treatment Attitude Strength in Study 2, Wave 1

Figure A2: HTEs on Attitude Strength by Belief Categorization (Descriptive/Evaluative)

Table A4: HTEs by Issue Categorization on Attitude Strength

	Attitude Strength (1-7)	
	Study 1	Study 2
Intercept	1.86*** (0.27)	2.40*** (0.29)
Focal Correction	-0.05+ (0.03)	-0.08** (0.03)
Distal Correction	0.00 (0.02)	-0.09** (0.03)
Hard Issue	0.00 (0.02)	-0.06+ (0.03)
Pre-Treatment Attitude Strength	0.62*** (0.05)	0.51*** (0.06)
Focal $\times$ Hard Issue	-0.15** (0.05)	-0.12+ (0.06)
Distal $\times$ Hard Issue	-0.12** (0.04)	-0.06 (0.06)
Num.Obs.	2485	1916
R2	0.357	0.257

+ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table A5: HTEs by Belief Categorization on Attitude Strength

	Attitude Strength (1-7)	
	Study 1	Study 2
Intercept	1.79*** (0.26)	2.30*** (0.28)
Focal Correction	-0.15*** (0.03)	-0.16*** (0.04)
Distal Correction	-0.04+ (0.02)	-0.15*** (0.04)
Beliefs: Descriptive-Evaluative	0.06+ (0.03)	-0.06 (0.04)
Beliefs: Evaluative-Descriptive	0.01 (0.03)	-0.02 (0.04)
Beliefs: Evaluative-Evaluative	0.04 (0.03)	0.04 (0.03)
Pre-Treatment Attitude Strength	0.63*** (0.05)	0.53*** (0.06)
Focal $\times$ Descriptive-Evaluative	-0.04 (0.07)	0.13+ (0.08)
Distal $\times$ Descriptive-Evaluative	-0.03 (0.05)	0.22** (0.07)
Focal $\times$ Evaluative-Descriptive	0.13+ (0.07)	0.04 (0.08)
Distal $\times$ Evaluative-Descriptive	-0.07 (0.06)	0.03 (0.08)
Focal $\times$ Evaluative-Evaluative	0.10+ (0.06)	0.04 (0.06)
Distal $\times$ Evaluative-Evaluative	0.07 (0.05)	-0.02 (0.08)
Num.Obs.	2486	1916
R2	0.353	0.251

+ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$