

An Experimental Comparison of AI-Enabled Semi-Structured Interviews and Fixed Surveys: Response Patterns, Quality and Representation*

Amanda Austin[†]

Edward Hohe[‡]

Ryan Kennedy[§]

Leib Litman[¶]

William Minozzi^{||}

Laura Moses^{**}

June 13, 2025

*Prepared for presentation at the 2025 annual meeting of the American Association for Public Opinion Research (AAPOR), St. Louis, MO. Authors are presented in alphabetical order.

[†]University of Houston

[‡]The Ohio State University

[§]The Ohio State University, Corresponding Author, kennedy.310@osu.edu

[¶]CloudResearch

^{||}The Ohio State University

^{**}The Ohio State University

Abstract

Artificial intelligence (AI) interviewing offers the potential to enhance the depth and quality of survey research, yet it may also introduce unintended consequences for the measurement of political attitudes. From the standpoint of respondents, more extensive interviews impose greater cognitive and temporal burdens, which may reduce engagement or increase satisficing. From the perspective of political behavior researchers, deeper probing may shift the subset of beliefs or arguments respondents access when forming opinions, thereby altering the distribution and apparent stability of measured attitudes. To investigate these dynamics, we conduct a large-scale, fully randomized experiment. We find that AI-mediated interviewing significantly increases both the volume and specificity of respondents' articulated rationales. While interview duration increases, we detect no decline in participant satisfaction. Importantly, however, AI interviews appear to shape subsequent responses: after offering justifications through AI interaction, respondents exhibit greater polarization in follow-up items. This result suggests that AI interviewing may alter the cognitive architecture of opinion formation by introducing a structured, reflective context that affects how individuals access and prioritize political considerations. These findings underscore the need for political scientists to critically assess AI interviewing not only as a measurement tool but also as a potential source of attitudinal change.

Keywords: artificial intelligence (AI), qualitative interviewing, public opinion, experiments

“Since the beginning of social research, students have tried to combine the value of detailed qualitative applications with the advantages of more formalized techniques which could be managed on a mass basis.” (Lazarsfeld, 1944, 59)

The incorporation of AI in public opinion research is widely regarded as transformational, particularly for enabling in-depth interviews at scale. By providing structured exploration of the cognitive processes underlying attitudes, AI offers researchers the means to better understand the foundations of political opinions. AI interviewers thus have the potential to advance our understanding of opinion formation and reasoning in mass publics (e.g., Xiao et al., 2020; Chopra and Haaland, 2023; Wuttke et al., 2024). Already, various AI interviewing approaches are being evaluated (Barari et al., 2025), and AI interviewers have been integrated into experimental designs (Velez and Liu, 2025) and used to develop adaptive Likert-style questions (Velez and Liu, 2025).

While AI interviewing promises to enhance the depth and quality of survey responses, it may have unintended consequences. Practically, more in-depth interviews demand greater time and cognitive effort from participants, resulting in substantially longer, more costly responses (Geiecke and Jaravel, 2024). Such demands may impact satisfaction and completion rates, potentially introducing new biases in selection and attrition.

AI interviewing, moreover, may destabilize both attitude measurement and response interpretation. In-depth interviewing often prompts participants to provide reasons, perhaps even transforming their opinions in the process, not unlike public deliberation (Mutz, 2008; Kalla and Broockman, 2020; Minozzi et al., 2024; Rawls, 1997; Thompson, 2008). As such, this new interview mode itself may constitute a stimulus, altering the accessibility and salience of considerations. Although recent evidence suggests reason-giving can have limited effects in some settings (Blumenau, 2025), the interactive and adaptive nature of AI interviewing may foster more deliberative processes than traditional reason-giving (Geiecke and Jaravel, 2024). AI interviewing thus puts distance between traditional top-of-the-head responses to closed-ended questions and summary responses preceded by AI interviews, leaving researchers with no clear answer about which constitutes a “true” attitude.

We address these questions through a large-scale survey experiment ($N = 2,243$) manipulating the presence and timing of an AI interviewer. We evaluate effects on reasoning, response consistency, response quality, item nonresponse (rolloff), satisfaction, and duration. We document that AI interviewing does substantially increase both the quantity and depth of respondents' reason-giving, and that these effects come at the cost of additional completion time, as expected. Despite these additional costs, AI interviewing had negligible effects on participant engagement and satisfaction. But AI interviewing also altered responses. Responses to follow-up (closed-ended) summary questions were more polarized for participants who responded first to a closed-ended question and then explained their reasoning in an AI interview. AI interviewing itself may therefore actively shape cognitive processing.

How Might AI Alter Survey Responses?

The emergence of AI-driven survey tools prompts the question of how and why AI interviewing might outperform or fall short of traditional data collection methods, with the potential to both improve and challenge current survey research practices. We develop several research questions to explore these possibilities.

From a respondent's perspective, participants are likely to spend more time interacting with an AI than they would with a standard set of closed-ended questions. As a result, respondents to AI interviews may find the experience less satisfying compared to more straightforward, less time-consuming survey formats, especially if they are paid survey-takers on a crowdsourcing platform. On the other hand, survey respondents may value the opportunity to express their opinions in greater detail, no longer constrained by predefined response options. AI has proven effective in sparking meaningful conversations (Croes and Antheunis, 2020; Yin et al., 2024), and studies suggest that people enjoy explaining *why* they behave or think a certain way, driven by a desire to be heard and understood (Hart et al., 2021; Hardin and Higgins, 1996; Brooks and John, 2018). These potentially opposing possibilities

led us to our first research question:

RQ1. Will respondents be more or less satisfied with the survey experience if it includes an AI interaction?

A second key question is whether the inclusion of AI interviews significantly influences the opinions expressed by respondents. The most widely accepted question-response model in public opinion research identifies four stages in a response: comprehension, retrieval, judgment, and response (Tourangeau et al., 2000; Willis, 2004). Comprehension refers to how the respondent interprets the question; retrieval is the process of recalling relevant information from memory; judgment involves evaluating that information and deciding how to answer; and response is the act of selecting an answer based on the available options, influenced by social desirability, question format, and the survey context. Similarly, Zaller (1992) introduces the concept of “sampling” in opinion expression, which aligns with retrieval and judgment, suggesting that individuals base their responses on the most recent and accessible considerations in their memory.

The potential impact of AI interviews on responses may be most apparent in terms of an ordering effect. There are several reasons to believe that the inclusion and timing of AI interviews could influence answers to closed-ended questions. For one, interacting with an AI interviewer may alter the cognitive processes involved in answering. If the interview begins with an AI interviewer, it may trigger fundamentally different thought patterns than those involved in simple retrieval. The process of thoughtfully articulating and justifying one’s reasoning may influence the way respondents ultimately answer. Some scholars have suggested that this “deliberation within” can significantly shape the opinions individuals express (Goodin, 2000; Goodin and Niemeyer, 2003). As such, leading with an AI interviewer could produce different responses than immediately starting with a closed-ended question.

RQ2. Will responses to closed-ended questions vary if they are preceded or followed by an AI interview?

Following up closed-ended questions with open-ended probes is not a new practice. In fact, much of the potential of AI interviews appears to build on what can already be achieved with fixed, non-responsive open-ended follow-up questions. Therefore, an important question is whether, and to what extent, AI interviewers can provide responses that are different from those offered by traditional methods. Beyond simple measures like word count, AI interviewers have the ability to probe more deeply into respondents' reasoning, as their follow-up questions can be tailored to the specific answer provided. This customization may allow AI to gather more detailed insights into the reasons behind a respondent's answers. Additionally, AI may be capable of identifying and addressing any misinterpretations of questions, offering respondents a chance to refine their answers. If AI interviews can yield deeper insight into respondents' thought processes—both in content and reasoning—they could substantially enrich the quality of survey data.

RQ3. How do responses to AI interviewers differ from those to standard, fixed follow-up open-ended question?

The cause-and-effect relationship suggested by RQ2 may also be reversed: starting with a closed-ended summary question could trigger different behavior in a follow-up AI interview compared to beginning with the AI interview itself. Respondents who first answer closed-ended questions and then engage in an AI interview might approach the interaction differently from those who start with an AI interview. People often rationalize their previously expressed opinions, retrieving information mainly to confirm their prior responses (Vedejová and Čavojová, 2022). While this heuristic generally applies to information search and acceptance, a similar process is possible in a communication context. In this case, the respondent may be more inclined to justify their closed-ended answer, rather than engage in a deeper exploration of their thoughts on the issue.

RQ4. Do responses to an AI interviewer depend on whether it is preceded or followed by a closed-ended summary question?

Interaction with an AI interviewer may also serve to solidify—or even polarize—attitudes, as individual survey responses are often unstable, even within the same survey, due to variations in the sampling process (Feldman, 1989). Engaging with an AI interviewer prior to answering a close-ended question might enhance the consistency of responses by reinforcing the information retrieved earlier, making it more likely that the sampling process will remain stable over time. On the other hand, if the AI interview follows an initial closed-ended question, it could prompt some level of internal deliberation, potentially leading to shifts in opinion. These changes could manifest when the closed-ended question is asked again later in the survey. The direction of this shift, however, is uncertain. Deliberation could moderate attitudes by bringing in counter-considerations more frequently, but it might also provide more room for motivated reasoning and confirmation bias, potentially leading to more polarized responses. This leads to two additional research questions:

RQ5. Will interaction with the AI interviewer alter response consistency?

RQ6. Will interaction with the AI interviewer polarize responses?

Finally, while these research questions are broadly framed, the effects they examine are likely to vary by context. Individual differences—such as personality traits, topic sensitivity, and engagement levels—may shape how respondents interact with an AI interviewer. For instance, traits like need for cognition (Cacioppo and Petty, 1982) may predict willingness to engage thoughtfully, while extroversion, conscientiousness, and openness (Herzberg and Brähler, 2006) could influence the depth of participation. Topic relevance, group identity, and attentiveness may further affect the nature of responses (Berinsky et al., 2014). Given these moderating factors, two key questions emerge about how AI interview effects vary across individuals and contexts:

RQ7. How do personality, demographic, and technology characteristics moderate the effects of the AI interviewer?

RQ8. How do standard attention checks match up with the quality of responses given to open-ended questions?

Data and Methods

To answer these research questions,¹ we used a fully randomized design to evaluate the effects of AI interviewing. We elected to conduct this research within the Engage platform designed by CloudResearch.² We opted to use this platform because: (1) it is one of the better-tested AI platforms, having been used throughout the 2024 election cycle in collaboration with the Sienna Poll (e.g., Siena College Research Institute, 2024); and (2) it features a robust design and deployment environment with multiple guardrails to ensure interview quality and data security.

At the time this study, Engage did not have the ability to randomize treatment assignment. To overcome this issue, we designed each treatment arm separately in Engage and embedded the Engage survey links into the HTML of a Qualtrics survey.³

Independent Variables

Our experiment had five treatment arms, based on the components outlined in Table 1.

¹We pre-registered all research questions and treatments at <https://doi.org/10.17605/OSF.IO/4NSB8> prior to fielding our study. See online appendix (sec. 6) for discussion.

²This experiment was reviewed by the IRB at (**omitted**) University: Study #2024E1130, 4 November 2024.

³The full survey text and HTML for Qualtrics integration is available in the online appendix (sec. 2).

Component	Question/Prompt
Closed-Ended Questions	
C1	How much do you agree or disagree with the following statement? There should be a way for undocumented immigrants to stay in the U.S. legally if certain requirements are met. [Strongly agree, Agree, Somewhat agree, Neither agree nor disagree, Somewhat disagree, Disagree, Strongly disagree]
C2	Overall, do you think increasing tariffs or fees on goods imported from trading partners will be good or bad for the United States? [Very good, Good, Somewhat good, Neither good nor bad, Somewhat bad, Bad, Very bad]
Preamble to Closed-Ended Question (Following Interview)	
P1	Now that you have discussed this and thought it through, please give us a summary of your thoughts on the issue.
Fixed Open-Ended Question	
F1	Why do you hold this position on this issue?
Open-Ended Opener for AI Interview	
O1	What do you think of the idea that undocumented immigrants should be allowed to stay in the U.S. legally if certain requirements are met?
O2	What do you think about increasing tariffs or fees on goods imported from trading partners to the United States?
Prompt for AI Interview	
AI	Ask follow-up questions to develop a fuller understanding of why the respondent does or does not think [issue], or which way they lean if they say they neither agree nor disagree. Probe for specific reasons they hold their opinions. Assess the strength of their opinion and if they are open to counter-arguments that might change their thinking on the policy. Listen for and probe on: Language or phrasing, Emotion, Tone, their thoughts and feelings regarding policies towards [issue]. Continue with the interview until you get a FULL picture of what this person was thinking. Ask for clarification or more details when responses are vague or general. Ask a maximum of three follow-up questions! Don't ask leading questions, let participants spontaneously tell you what they think without putting words in their mouths. Also, always ask only one question at a time!

Table 1: **Components used to construct treatments.** Table shows the components that are used in the construction of treatments in Table 2. AI portion shows the prompt used to generate the conversations, as illustrated in Figure 1.

Four of these treatment arms included two closed-ended items authored by Pew Research Center (2024, 2018) that gauge support for recent immigration and tariff policy. The control condition, **Closed Only (CO)**, asked only the closed-ended questions without further discussion. A second non-AI treatment, **Closed, then Fixed (CF)**, included the same closed-ended items followed by a generic open-ended prompt: “Why do you hold this position on this issue?” Although more sophisticated techniques exist for fixed follow-up questions, such generic prompts are common in applied research.

The other three treatment arms all incorporated AI at varying stages. The **Closed, then AI (CAI)** follow-up condition entailed an open-ended AI interview after the closed-ended questions. The **AI, then Closed (AIC)** follow-up arm reversed this order, starting with a general preamble asking what they think about the issue (O1 and O2), followed by the AI interview and closing with the closed-ended questions. Finally, we included an **AI Only (AIO)** condition, which featured only the introductory preambles, followed directly by the AI-conducted interviews. Figure 1 shows an example of what a treatment (CAI) looks like in a simulated interview on the Engage platform. Table 2 records how the components are combined to create the five treatment conditions.

Treatment	Composition
Closed Only (CO)	C1 and C2
Closed then Fixed Followup (CF)	(C1 + F1) and (C2 + F1)
Closed then AI Followup (CAI)	(C1 + AI) and (C2 + AI)
AI then Closed Followup (AIC)	(O1 + AI + P1 + C1) and (O2 + AI + P1 + C2)
AI Only (AIO)	(O1 + AI) and (O2 + AI)

Table 2: **Composition of treatments.** Table shows the composition of treatments using keys from Table 1.

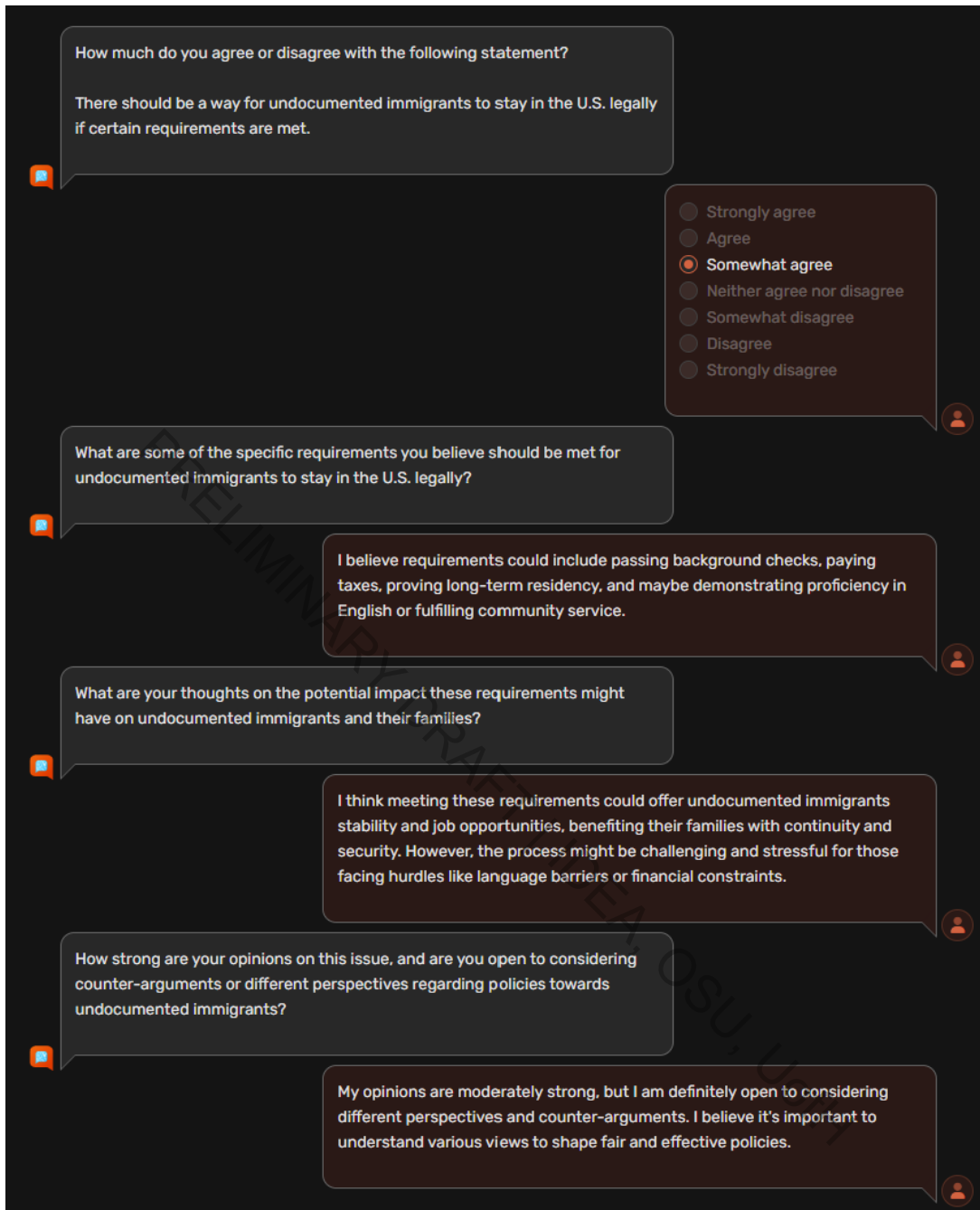


Figure 1: **Example (simulated) of CAI treatment arm in Engage platform.** Figure shows a simulated conversation between AI agents on the immigration question.

Dependent Variables

We measured multiple outcomes to systematically assess our research questions. First, we assessed respondent behavior in the survey itself, including the time respondents take to complete the survey (in seconds) and the total number of words used in their text responses. For four experimental arms (excluding AIO), we analyzed responses to closed-ended items, both when they were asked in the Engage interview (which we label the ‘Interview’ responses) and a second time after the demographic and personality questions (‘Post-Interview’ responses), allowing us to assess the consistency of respondents’ answers. We also measured polarization by calculating the distance of each response from the median option and then taking the difference in these distances, such that positive values indicate polarization, negative values indicate moderation, and zero indicates no change in response. Additionally, we asked all respondents to rate their satisfaction with the survey experience on a seven-point scale and to describe their experience with three open-ended questions asking: (1) what they liked about the survey, (2) what they disliked about the survey, and (3) if they had any additional comments for the researchers. We coded these open-ended questions manually to record sentiment towards the AI interaction, as well as what other parts of the survey the respondents liked or disliked. Finally, we also used AI tools to measure the content of the text responses in Engage, specifically in terms of the reasons given by respondents for their opinions on immigration and tariff policies.

Covariates

We measured a number of respondent characteristics, both to assess balance across experimental conditions and to identify differences in response to AI interviewing (RQ7). At the beginning of the survey, we included a standard attention check and recorded whether the participant answered correctly (Berinsky et al., 2014). We did not, however, remove these participants from the study because we are interested in how attentiveness affected responses (RQ8). We drew standard demographics from CloudResearch: age, education, gender, par-

tisanship, racial/ethnic identity, political interest, and political knowledge. Additionally, we included two standard psychometric batteries: the Big-5 personality index (short form), which measures extroversion, agreeableness, conscientiousness, neuroticism, and openness (Herzberg and Brähler, 2006), and a short form measure of respondents’ need for cognition, which may relate to their willingness to engage in thoughtful conversation on an issue (Cacioppo and Petty, 1982). We used equivalence testing (Hartman and Hidalgo, 2018) to assess balance on these covariates and found no imbalance across any pair of treatment conditions for any covariate.⁴

Recruitment

We recruited 2,243 respondents from CloudResearch’s Connect platform between 6 and 11 December 2024, who were randomly assigned into one of the five treatment arms. CloudResearch has been used by a large number of experimental studies, replicates experimental findings from other major survey sources, and provides better data than many alternative online providers (Douglas et al., 2023; Hauser et al., 2023).

We took advantage of the option for CloudResearch to match the age, gender, ethnicity and race characteristics of our sample against US Census values. However, to complete the survey in a timely fashion, we had to relax these quotas towards the end of data collection. This produced slight shortfalls in male, Hispanic/Latino, and white respondents, along with a larger shortfall in respondents over 60 years old. These quota fulfillment ratios are reported in the online appendix (sec. 3).

Each respondent was paid a flat rate of \$3 for their time, regardless of the treatment arm to which they were assigned. We estimated that the survey would take about 10-15 minutes for the most intensive treatment arm based on early pilots. The actual median response times ranged from about 6 minutes for the CO survey to about 14 minutes for the AIC survey. The final payment came out to an hourly rate of ranging between \$30 per hour to

⁴All equivalence tests were based on the “strict” null of a difference of 0.37SD, used a cutoff of $p = 0.05$ (two-tailed), and were corrected for false discovery rates.

\$12.86 an hour.

Analysis

Unless otherwise noted, we utilize basic Bayesian Ordinary Least Squares (OLS) regression with uninformative priors. We use this method primarily because of the ease of plotting the full distribution and deriving a variety of statistics using draws from the posterior distribution (Jackman, 2009). Distributions are—unless otherwise noted—derived from 4,000 draws from four MCMC chains. We rely on standard convergence checks to ensure the Markov chains sufficiently explore the posterior distribution.

Survey Cost, Satisfaction and Duration

A first question is whether—and to what extent—incorporating AI interactions entails additional costs. The monetary cost to conducting these interviews was minimal. Engage is currently free for researchers, but, even were it not, the total cost for conducting these interviews based on the number of tokens involved, and assuming a very inefficient system of implementation, would have been around \$0.008 per interview for a total of about \$10.44 for all the interviews.⁵ Put simply, the implementation costs of AI interviews is dwarfed by participant costs.

A further potential cost has to do with participant reactions to the interview process under the assumption that respondents demand more for tasks they like less. We analyzed self-reported satisfaction with experiences on our survey. Figure 2 (left panel) displays marginal effects for each treatment arm (i.e., the mean within the condition minus the grand mean). Each 95% interval includes the overall average. Satisfaction is generally higher in the control Closed Only (CO) condition and Closed then Fixed Followup (CF), which do not include AI interviewing, but is also similar in Closed then AI Followup (CAI), which

⁵Calculations based on pricing for Open AI’s GPT-4o on 7 March 2025. Details in online appendix (sec. 5).

did incorporate AI. Satisfaction is lower than average in both AI Only (AIO) and AI then Closed Followup (AIC), and some cross-treatment comparisons yield significant differences. To formally test the difference due to AI specifically, we estimate bivariate linear regression models of satisfaction on an indicator of whether treatment included an AI component. On this seven-point scale, the coefficient is negative, at -0.06 with 95% interval $[-0.11, 0.02]$ for the full sample.⁶ We conclude that there is—at most—evidence of a very small decline in satisfaction due to AI interviewing.

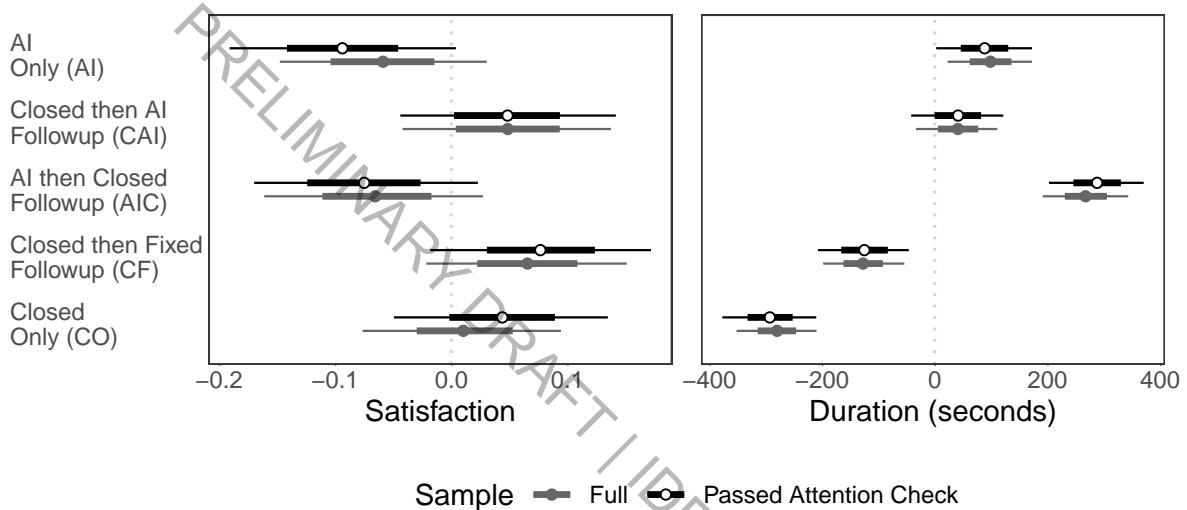


Figure 2: **Negligible loss in satisfaction, despite increased duration.** Satisfaction is measured on a 7-point scale; duration, in seconds. Points show the estimated marginal effects. Bars show the quantile intervals for the 67% and 95% credible intervals estimated from 4,000 draws from the posterior distribution.

The absence of satisfaction costs does not seem to be an artifact of differential attrition. Of the 2,253 participants we recruited to start the survey, only 10 failed to complete it, and all of these respondents rolled off before starting the Engage component of the survey, which includes all AI interactions. There was, however, a small level of differential completion of the text-entry components. Further inspection of this pattern revealed that it was almost certainly a technical issue dealing with when the “next” button was displayed.⁷

⁶Results for this comparison and all subsequent ones are similar if we drop those who fail the attention check.

⁷The online appendix (sec. 8.2) provides a full breakdown of this issue.

The negligible difference in satisfaction appears even though there were sizable (uncompensated) differences in survey duration (see Figure 2, right panel). As noted above, all respondents received \$3 regardless of treatment. Duration varied dramatically across treatments, ranging from a median under 7 minutes for the closed-ended-only control (CO) to 14.5 minutes for an AI-first interaction (AIC). The result was a median hourly rate of \$25.90 for CO, \$20.49 for CF, \$14.59 for CAI, \$13.90 for AIO, and \$12.41 for AIC.

These differences in duration stand in stark contrast to the negligible differences in satisfaction. One explanation is that these respondents were professionals who are keenly aware of pay scales and thus unsurprised and undisturbed by any exigencies of participation. But another explanation is that the experience with AI interviewer was sufficiently novel and interesting as to be intrinsically motivating.

Some evidence for the latter explanation emerges from qualitative data. Respondents were asked at the end of the survey what they liked best, what they disliked most, and if they had any additional comments for the researchers. We analyzed these comments manually. Despite no requirements for response, 1,057 of 1,422 respondents who interacted with an AI interviewer left such a comment (74%). Of those, 470 (44%) are positive, while only 126 (12%) leave some type of criticism about the AI, and of those who leave a negative comment, 47 (37%) leaven their criticism with a positive comment about the chatbot, e.g., “I wish the chat with the chatbot could have gone on a little longer.” Overall, for every negative comment about the AI interviewer, there were six positive comments.⁸ While novelty might depress survey enjoyment by increasing cognitive demands it might also elevate the survey experience by breaking up the monotony of what respondents called “bubble hell”. These qualitative data support the latter. Those that mentioned the survey was new, unique, or novel did so exclusively positively and one noted that this was their favorite AI survey suggesting the satisfaction benefits of novelty endure at least beyond the first couple experiences.

We conclude that there was little indication that incorporating an AI interview into the

⁸A more complete breakdown of the negative comments is available in the online appendix (sec. 8.3).

survey resulted in lower survey satisfaction, even when it decreased hourly pay.⁹ Indeed, this evidence is consistent with the possibility that many respondents experience intrinsic *enjoyment* because of the inclusion of the AI interviewer.

Word Count and Reason Provision

Perhaps the most compelling questions arising from this research are what unique insights interviews can reveal that cannot be uncovered through a simple open-ended follow-up—such as asking respondents why they answered as they did—and how the integration of AI into the interview process influences these outcomes.

To begin, respondents in the AI interview treatment arms have a substantially higher word count than those who received the fixed open-ended question (CF) (RQ3). Despite the AI interviewer being instructed to ask no more than three follow-up questions,¹⁰ in conditions with the AI interviewer, respondents use 70 [64, 75] more words answering immigration questions and 57 [50, 64] more words answering tariff questions. These rates more than double the word counts in CF.

RQ4 asks about the specific difference caused by following up with an AI interviewer, as opposed to beginning the experience with such an encounter. While we see no such order effect for immigration, we do observe that the AI follow-up (CAI) produces significantly fewer words than the other two AI conditions on tariffs. This finding suggests that, at least in some circumstances, the procedure of interviewing prior to asking a closed-ended question encourages more detailed responses from the respondents.

Moving beyond simple word counts, we analyze whether there were differences in the reasoning provided by respondents. LLMs have proven highly proficient in text summarization (Mathis et al., 2024; Hake et al., 2024) and annotation (Savelka and Ashley, 2023;

⁹Even the lowest pay amounts in our survey, however, are still within the bounds of what is usually considered “good” for an online survey. Prolific provides a compensation rating system (<https://www.prolific.com/>). \$12 an hour is considered “good” compensation in this rating. Our median compensation was above this threshold for all treatment arms.

¹⁰An instruction which, admittedly, the AI does not always follow; see below.

Törnberg, 2023; Wang et al., 2023). We leveraged these capabilities, in conjunction with human oversight and review, to analyze specific parts of the output of these interviews and the degree to which they are influenced by the use of AI interviewing.¹¹ We used an LLM to summarize the reasons given for respondents’ positions on the immigration and tariffs issues, splitting responses into groups that support, oppose, or are neutral on each policy based on closed-ended responses for each treatment.

LLM output is stochastic, yielding different outputs on repeated runs. Therefore, we took a two-step approach to coding. First, for each combination of issue and position, we asked the LLM to summarize the “major reasons” provided in the interviews, along with a summary label of their prevalence.¹² The results produced an approximately normally distributed count of topics generated, with a median of 7 and a mean of 7.3. Among runs, the maximum number of topics generated across all conditions is 16 and the minimum is 1. Table 3 shows an example of this output for those who support allowing undocumented workers to stay in the US.¹³

Our second step was to draw a random set of labels from those summaries for each policy and position within the CAI treatment. We then asked the LLM to label comments using these categories.¹⁴ We review of a random sample of 300 labels from each group revealed the labeling to be quite accurate, at least within the tolerance expected from a skilled research assistant.¹⁵

While the topic summaries of the LLM were quite good, the estimates of the prevalence of different reasoning in Table 3 were not. The rightmost column of the table reports the prevalence levels actually observed (for the CAI treatment). In addition, there appears to be at least one redundant category. “Security vetting” was not used at all in labeling the

¹¹We used the OpenAI API for all these analyses. None of the data was stored or used for training by OpenAI, see online appendix (sec. 8.5).

¹²Details in the online appendix (sec. 9.1).

¹³The illustrated example is simulation 35 from the CAI treatment and is randomly drawn from the simulations for that treatment.

¹⁴Details in the online appendix (sec. 9.2).

¹⁵We found only 4 cases (1.3%) where the label was clearly incorrect and another 16 (5.3%) where it was unclear if the label was correct based on the pulled quotation. Details in online appendix (sec. 9.2).

Reason	Summary	Labeled Prevalence	Actual Prevalence
Economic contribution	Many participants highlighted that undocumented immigrants fill essential jobs that U.S. citizens are unwilling to do, thus contributing significantly to the economy.	high	0.437
Family and humanitarian concerns	Participants expressed concern over family separation and suggested that individuals who have built lives and families in the U.S. should have a path to legality.	high	0.135
Cultural and workforce diversity	Participants acknowledge the role of immigrants in enriching U.S. culture and workforce diversity, adding that immigrants rejuvenate local communities.	medium	0.067
Long-term residents	Participants argued that undocumented immigrants who have lived in the U.S. for many years and are law-abiding should be allowed to stay legally.	medium	0.141
Lack of criminal record	Participants frequently cited the absence of a criminal record as a crucial condition for allowing undocumented immigrants to stay legally.	medium	0.486
Moral and ethical considerations	Some participants talked about moral obligations and empathy for individuals fleeing from dire conditions, suggesting they should be treated humanely.	medium	0.303
Legalization process improvements	Participants proposed simplifying and accelerating the legal process for undocumented immigrants as current systems are seen as burdensome and slow.	low	0.138
Security vetting	A few participants indicated the need for background checks and vetting as reasonable measures to ensure the safety and integrity of the immigration process.	low	n/a

Table 3: **Example of LLM summary of reasons given for support for allowing undocumented workers to stay in the US.** The treatment for this summary is the closed-ended question followed by an AI interview (CAI).

responses, likely because it was redundant with the “lack of criminal record category.” While the labeled prevalence measures from the first step of our reason coding sometimes aligned with actual prevalence (e.g., “Economic contribution” or “Moral and ethical consideration”), the relationship seems otherwise scattered. Thus, while the LLM proved effective at summarizing the reasons themselves, it does not do a good job of estimating their prominence or of avoiding non-mutually-exclusive categories.

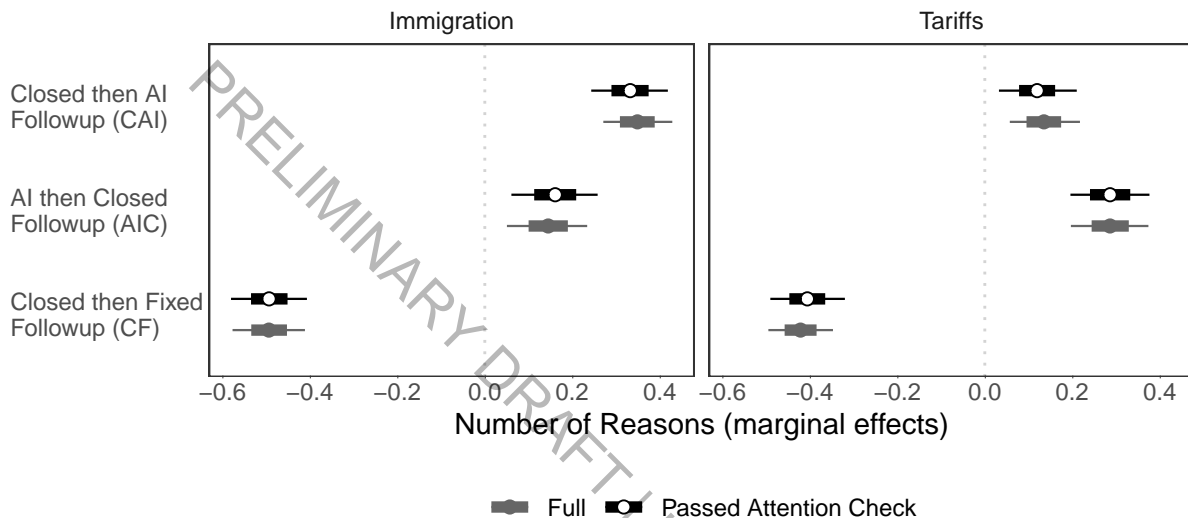


Figure 3: **Interaction with the AI Interviewer Boosted Reason Provision.** Points show marginal effects, and bars show 67% and 95% intervals.

We can now address our research questions RQ3 and RQ4 using this measurement of reason-provision. First, we counted the number of reasons each respondent provided separately by issue and regressed the results on treatment indicators (see Figure 3). The results are clear: the AI interviewer not only encouraged respondents to write more text, it caused them to increase the number of reasons they offered. A bivariate regression on an indicator for the AI interviewer reveals that it increases the number of reasons by 0.75 [0.65, 0.85] for immigration and 0.63 [0.53, 0.72] for tariffs for the full sample.¹⁶ However, we see no clear order effect of where the AI was placed, as the results for the two issues have opposing signs (whose 95% intervals exclude zero). We conclude that AI interviewers were sparking longer,

¹⁶Results are very similar when dropping those who fail attention checks.

more varied conversations than what is achieved by a fixed prompt.

Finally, we estimate separate regression models of binary indicators for whether a respondent provides a reason on treatment condition indicators, excluding control (CO, which lacks text output) and AI Only (AIO, which lacks a closed-ended interview question, and so cannot be binned into support/neutral/oppose positions). Figure 4 presents the results.¹⁷

First, we see that in most cases, the fixed follow-up question (CF) yields a generally lower propensity to provide reasons. Pooling across all reasons, both the AIC and CAI conditions have provision rates of 24%, while that for CF is only 14%. Combined with the above evidence on word counts, the answer to RQ3 appears to be that the AI interviewer both lengthens and broadens engagement.

This difference in engagement also depended on whether the AI interviewer precedes or follows up a closed-ended question (RQ4). At times, this difference was dramatic. Across all 41 reasons (including the neutral positions), the 95% interval for the difference between **CAI** and **AIC** excludes 0 in 22% of cases. Those differences are not univalent—follow-up with the AI interviewer is more likely to cause those who support tariffs to invoke domestic production, while initiating the experience with the AI interviewer is more likely to yield a discussion of job protection. The implication is that the placement of the AI interviewer—whether before or after a closed-ended question—can meaningfully shape the content and depth of the responses, demonstrating how design choices in interview sequencing can influence the substance of the conversation.

The inclusion of AI interviewing has a substantial impact on the topics discussed and the reasoning patterns detected for popular political positions. The interviewing process suggests that these arguments would not emerge by simply asking a follow-up question—a method commonly used in current studies, when open-ended questions are included at all. Experimental studies are required to uncover the extent to which the arguments preferred are sufficiently important for the extra time and cost of AI interviews.

¹⁷For space reasons, we omit the neutral position, which has only 161 respondents for tariffs and 66 for immigration.

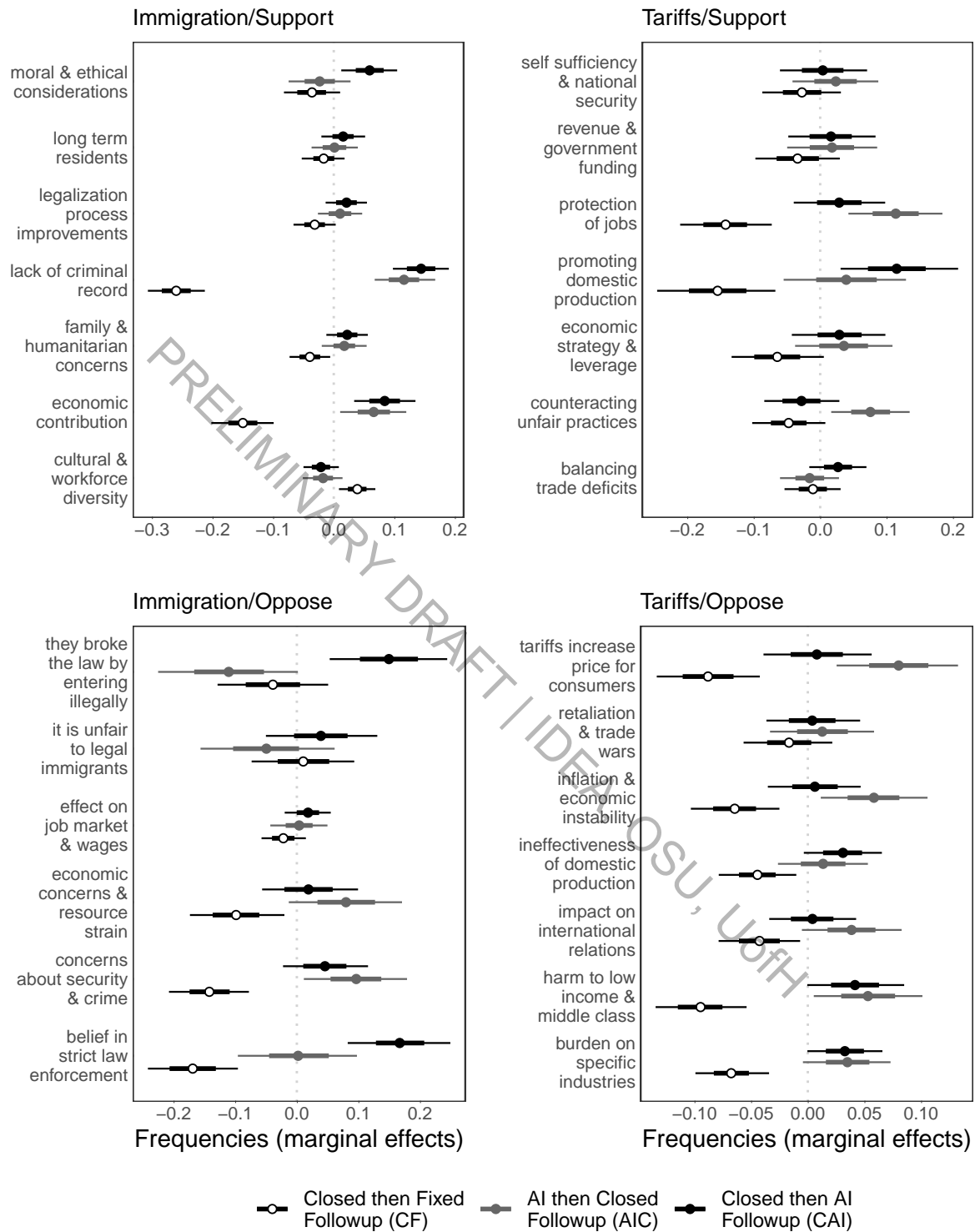


Figure 4: **Interaction with the AI Interviewer Substantively Altered Reason Provision.** Points show marginal effects, and bars show 67% and 95% intervals.

Effects on Closed-Ended Responses

Beyond participant experience and detail of responses, there are fundamental questions about whether and how AI interviewers might alter closed-ended responses. To shed light on these questions, we embedded the closed-ended items into the Engage interview. We designate these questions as the ‘Interview’ closed-ended items because they are asked as part of the broader interview in Engage. Treatment arms vary based on whether these questions come before (CAI) or after (AIC) interacting with the AI interviewer, whether there is only a traditional fixed open-ended follow-up (CF), or whether there is no follow-up at all (CO). We can therefore probe for effects of the presence and timing of the AI interviewer by analyzing responses to these closed-ended questions across all four treatment arms. The key contrast is between AIC and the other three treatment arms (RQ2).

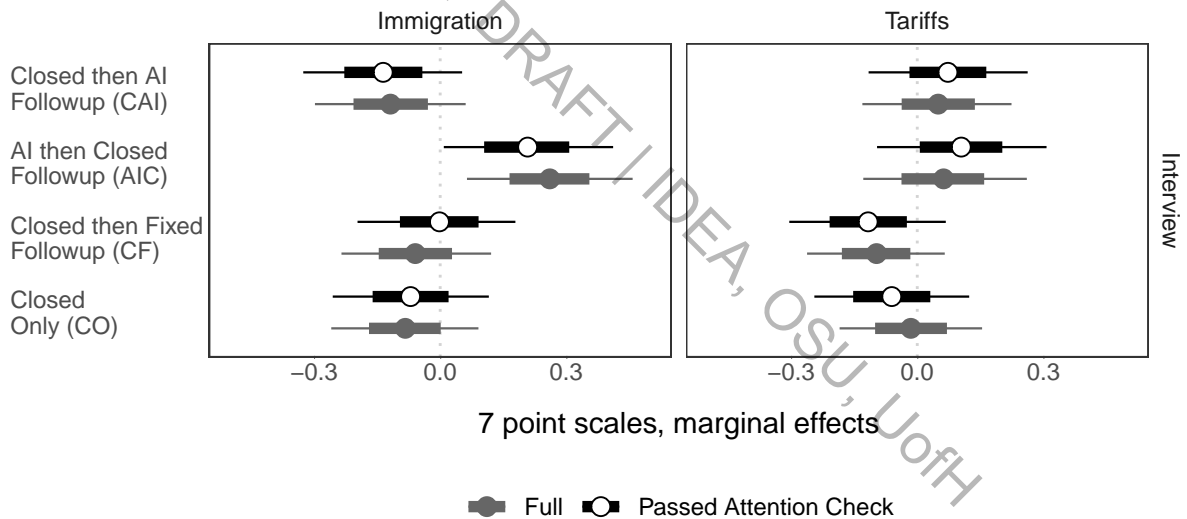


Figure 5: **Interacting with the AI Interviewer Sometimes Alters Responses to Closed-Ended Questions.** Points show marginal effects, and bars show 67% and 95% intervals.

Results appear in Figure 5. For immigration, interacting with the AI interviewer shifts responses slightly toward greater support. A bivariate regression of responses to this question on an indicator for treatment arm AIC yields a coefficient of 0.35 [0.13, 0.57]. In contrast, we see no such effect for tariffs. These findings suggest that interacting with the AI interviewer

can move responses, but that such effects are likely to be context-specific.

After respondents finished the interview on the Engage platform, the survey continued on Qualtrics. After a series of personality and political question, we fielded the same closed-ended items again. We use responses to these ‘Post-Interview’ questions to gauge the extent to which the AI interviewer provokes change (RQ5). To do this, we first model a binary indicator capturing any change between Interview and Post-Interview responses, and second, we analyze the magnitude of change by measuring the absolute difference between the two.

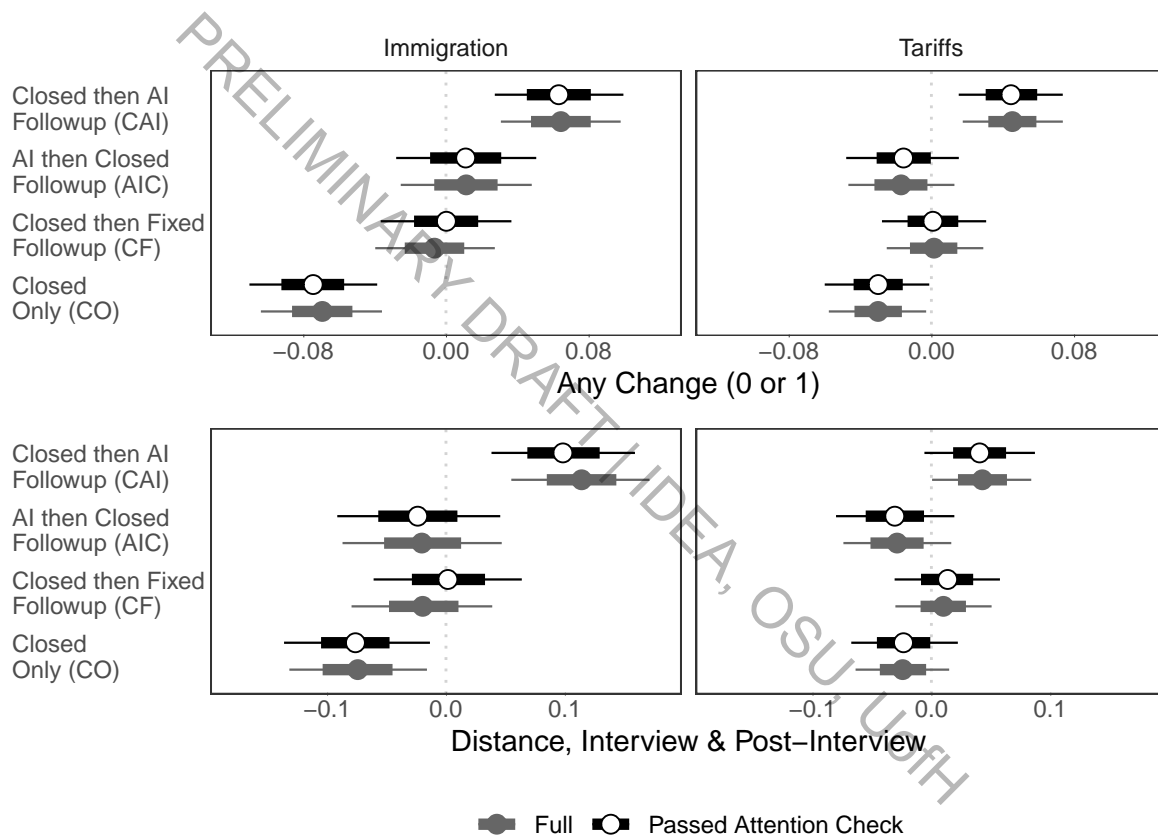


Figure 6: **Followup with the AI Interviewer Changed Responses.** Points show marginal effects, and bars show 67% and 95% intervals.

Figure 6 displays the results. Across both issues, we see evidence that interacting with the AI interviewer *after* providing a response to the closed-ended question (CAI) causes some respondents to change. There is inconsistent evidence that *any* follow-up probe for explanation causes such changes—the difference between “any change” rates for CF (the

fixed-text probe) and CO (control) is 0.06 [0.02, 0.11] for immigration and 0.03 [−0.01, 0.07] for tariffs. But the corresponding contrasts between CAI and CO is 0.13 [0.09, 0.18] and 0.08 [0.04, 0.11]. Moreover, even though collapsing all changes into this binary measure entails eliminating any information about the size of these changes, this pattern persists in the analysis of the distance between Interview and Post-Interview responses.

Given that the AI interviewer provokes changes, we next analyzed the extent to which participants select the most extreme response options (RQ6). To do so, we calculate the distance of responses to the scale median, then subtract the distance observed during the Interview from that during Post-Interview, and then regress that quantity on treatment arm indicators.

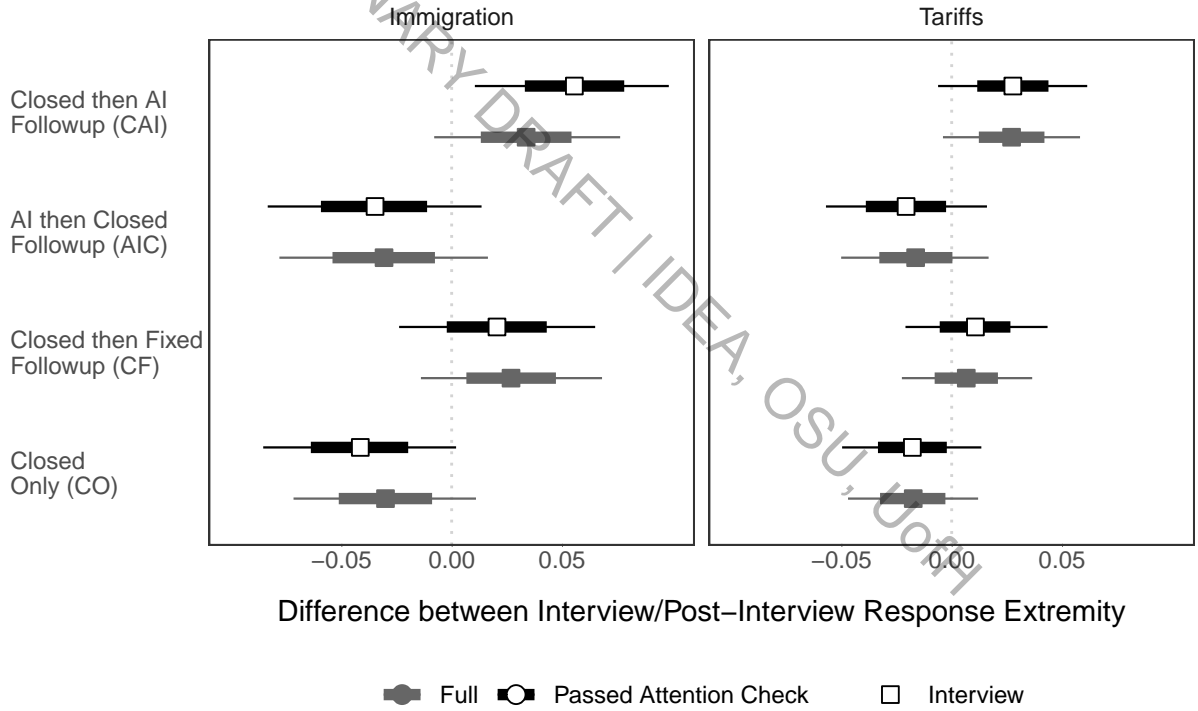


Figure 7: **Followup Probes Polarized Responses.** We calculate the outcome variable by taking the distance of responses to the scale median, and subtracting the distance observed during the from that during Post-Interview. Points show marginal effects, and bars show 67% and 95% intervals.

Results are depicted in Figure 7. We observe that—relative to control (CO)—follow-up

with the AI interviewer (CAI) polarizes responses on both issues. For immigration, the effect was 0.064 [0.007, 0.123], and for tariffs, 0.045 [0.02, 0.090]. This effect is not limited to the AI follow-up, however, as we see similar changes for follow-up with a fixed open-ended prompt (CF, 0.057 [−0.003, 0.114] for immigration and 0.024 [−0.018, 0.066] for tariffs). These effects are of small magnitude. The outcome variable has a theoretical range from −3 to 3, and, in keeping with the modest change rates visible in Figure 6, most of its values are zero. Nevertheless, these analyses reveal that, when follow-up provokes change, it tends toward response polarization.

This evidence suggests AI interviewers incite several changes in responses. First, conducting a preliminary AI interview impacts responses on immediate follow-up closed-ended questions. Second, reversing that order, following up a closed-ended item with a probing AI interview destabilizes responses, even beyond changes caused by fixed follow-up prompts. Finally, while these changes are relatively uncommon, they tended to push toward extreme response options. These results suggest some level of “deliberation within” may be triggered by the use of AI interviewing, altering the cognitive process used by respondents when formulating responses to closed-ended questions and, perhaps, crystallizing attitudes.

No Clear Evidence of Heterogeneity

The effects described in the previous two sections are full sample averages—with one important exception: we find no evidence of differences depending on whether respondents pass the attention check. Thus, the answer to RQ8 is that standard attention checks may have limited value for quality control in AI interviews.

Beyond the attention check, RQ7 asks more generally about heterogeneity across treatment conditions depending on covariates. To probe for such heterogeneity, we re-estimate variants of all bivariate regressions reported above, now including multiplicative interactions with each covariate listed above.¹⁸ We perform these estimations for both the full sample

¹⁸Covariates with more than three values were coded into terciles.

and for the subset of respondents who pass the attention check. The result is 588 different comparisons.

With such a diffuse question and so many comparisons, we would ordinarily eschew examining whether and to what extent we would observe significant coefficients. But in this case, we do so because of how little evidence of structured heterogeneity we observe. In only 22 cases (about 3.7%) do we observe a coefficient whose 95% interval excluded zero. Fewer still are robust to dropping those who failed the attention check: only 16 (i.e., 8 pairs) of the 22 cases. These estimates include the effect of the AI interviewer duration for both Political Knowledge (95 [11, 182], all coefficients for the full sample), and the indicator for Black¹⁹ (350 [2, 683]); the order effect on the Interview immigration closed-ended question for Extroversion (-0.33 [$-0.59, -0.07$]), Openness (-0.27 [$-0.53, -0.005$]), and Democrat (0.23 [0.005, 0.45]); the same effect for tariffs for Democrat (-0.26 [$-0.48, -0.05$]); the order polarization effect for Internal Efficacy (-0.05 [$-0.10, -0.01$]); and the effect of the AI interviewer on word count for Political Knowledge (8.7 [1.5, 15.9]).

We do see weak evidence of consistency here. The common effect for Democrats on the closed-ended order effects suggests the possibility that out-party members may be more susceptible to engagement with the AI.²⁰ Further, the effects of Political Knowledge on duration and word count suggest that the most knowledgeable will similarly be more apt to engage more deeply, likely due to information resources in long-term memory. The broader takeaway, however, is that there is little evidence that the effects reported above are driven primarily by any specific subgroup of respondents.

¹⁹Note that this comparison is with the omitted racial/ethnic category (“other race/ethnicity”); this coefficient is similar to those for Asian, Latino, and White.

²⁰At the time this study was in the field, Democrats had lost the election but the new government had not yet taken power.

Analysis of the AI Interviewer

We now turn to the AI interviewer, evaluating its behavior to identify considerations for scholars conducting similar research and highlight avenues for future study. Although all our AI interviews use identical prompts, the stochastic nature of large language models—and the unique context of each interview—can lead to variation in performance.

Violations of Prompt

As a first step, we used an LLM to evaluate whether the AI interviewer complied with their instructions.²¹ In all AI treatment arms there was evidence that AI interviewers deviated from their prompts. About half of all interviews, violated the rules laid out in the prompt in some manner.

On face, this suggests substantial issues with using AI interviewing. There are, however, at least two types of violations. The first type we label *formal violations*. This is where the AI interviewer departs from a structural instruction. These types of violations are not necessarily severe. An AI asking extra questions of an especially uninformative respondent that stretches beyond the 3 follow-up limit we placed on it might be annoying to the recalcitrant respondent, but it is the type of behavior we would also expect from a human interviewer. Similarly, asking double-barreled questions when the AI attempts to cover all the areas we requested within a constrained number of questions is not necessarily harmful in conversational interviews, provided the interviewee adapts and answers both parts of the question (Mukherjee, 2023). In addition, these types of violations can usually be addressed by small changes in the survey and prompt structure (Atreja et al., 2024). If, for example, the AI is asking double-barreled questions to cover all the topics in a small number of questions, this suggests that either the topics should be split up or that the question limit should be relaxed.

The second type of violation is what we label *biasing violations*. This is where the LLM

²¹Details of this evaluation are available in the online appendix.

violates rules about remaining neutral in its questioning, becomes hostile in its questioning, and/or goes off-topic. These types of violations have the potential to result in unusable data and dissatisfied respondents.

An overwhelming number of interviews that did not comply with the original AI prompt committed formal violations, specifically, they had double-barreled questions. Double-barreled questions are the violation of compliance in more than 95% of cases in both interviews about immigration and tariffs. Given the prevalence of these interviewer deviations from interview instructions, we then analyzed the extent to which participants answered both parts of the double-barreled questions posted by the AI interviewer. To do so, we again asked the LLM to evaluate the AI-conducted interviews.²² Where there were double-barreled questions, a little over 60% of respondents answered both parts of the questions.

Then, we look for evidence of biasing violations. Specifically, we consider whether the AI interviewer corrected respondents or “pushed back” against opinions or stances shared by respondents over the course of the interview. We again rely on an LLM to identify these behaviors in the interviews.²³ This process identified AI corrections in approximately 6% of the interviews and AI pushback in about 5%. A closer examination of the flagged instances suggests that even these figures may overestimate the prevalence of such behaviors. Many cases categorized as pushback involved the AI interviewer requesting examples, elaborating on potential respondent positions, or providing additional details about the conversation topics. Overall, we found very little indication of inappropriate interviewer behavior. This was reflected in the comments from participants. Four participants complained of bias and/or pushback in the interviews. In all of these cases, we could not find specific indicators for this evaluation by the respondent.²⁴

²²Details in online appendix (sec. 7).

²³Details in the online appendix (sec. 7.1).

²⁴And, in one case, the respondent was clearly trying to mess with the AI.

Interviewer Objectives

Next, we considered the extent to which AI interviewers are able to both deepen and expand discussion. Here again, we used LLM labeling with human evaluation of coding.²⁵ In most interviews, we found that the AI expanded the discussion ($\sim 70\%$). In almost all of the interviews, about 99%, we saw that the AI deepened discussion. These results suggest that AI interviewers are acting like skilled human interviewers, probing for reasons and encouraging elaboration from participants.

Interviewer Consistency

Finally, we consider the issue of reproducibility in AI interviewing. An AI interviewer is able to probe and ask substantive follow-up questions that are individualized to the interview. However, there is a balance to be achieved—we do not want AI interviews to be completely different for each participant, undermining comparability and reproducibility.

We evaluate differences in questions asked by the AI interviewers using Kullback-Leibler (KL) divergence to quantify the differences between respondents within the three AI interviewer treatment groups. KL divergence measures how much these distributions of words diverge between interviews.

The KL divergence in the AI interviews treatments were all around 0.81. Thus, in all of the AI interviewers the language used in the course of the conversation is similar, well-aligned, and consistent across topics and treatment conditions.²⁶

Conclusions

We present a large-scale randomized survey experiment varying both the presence and timing of an AI interviewer to investigate its effects across multiple dimensions. A central finding

²⁵Details in online appendix (sec. 7.2).

²⁶Details in online appendix (sec. 7.3).

is that AI interviewing does not negatively affect item nonresponse (rolloff) or participant satisfaction, despite increasing interview duration and cognitive engagement. Notably, AI interviewing substantially enhances both the quantity and depth of respondents' articulated reasoning, suggesting improved access to the underlying, internally reflective deliberation informing public opinion. However, the AI interviewer elicits different responses than traditional methods. While fostering more explicit reasoning, this increased engagement may induce cognitive shifts, evidenced by changes in subsequent closed-ended responses, including signs of opinion polarization. In this sense, AI may act as an intervening stimulus—amplifying the salience of particular considerations and prompting a more deliberative reasoning process.

Further research should explore the cognitive mechanisms through which AI-mediated interviews influence the salience and weighting of considerations during opinion formation and assess the extent to which these effects generalize across diverse issues and populations. Additionally, researchers should examine how AI interviewer characteristics—such as affect, tone, and conversational structure—impact respondent behavior, reasoning, and overall interview quality. Understanding these dynamics is critical for evaluating the long-term implications of AI interviewing on the reliability, stability, and comparability of public opinion measures across time and contexts.

References

- Atreja, S., Ashkinaze, J., Li, L., Mendelsohn, J., and Hemphill, L. (2024). Prompt design matters for computational social science tasks but in unpredictable ways.
- Barari, S., Angbazo, J., Wang, N., Christian, L. M., Dean, E., Slowinski, Z., and Sepulvado, B. (2025). Ai-assisted conversational interviewing: Effects on data quality and user experience. *arXiv preprint arXiv:2504.13908*.
- Berinsky, A. J., Margolis, M. F., and Sances, M. W. (2014). Separating the shirkers from the workers? making sure respondents pay attention on self-administered surveys. *American journal of political science*, 58(3):739–753.
- Blumenau, J. (2025). Does reason-giving affect political attitudes? *British Journal of Political Science*, 55:1–21.
- Brooks, A. W. and John, L. K. (2018). The surprising power of questions. *Harvard business review*, 96(3):60–67.
- Cacioppo, J. T. and Petty, R. E. (1982). The need for cognition. *Journal of personality and social psychology*, 42(1):116.
- Chopra, F. and Haaland, I. (2023). Conducting qualitative interviews with ai.
- Croes, E. A. and Antheunis, M. L. (2020). 36 questions to loving a chatbot: are people willing to self-disclose to a chatbot? In *International Workshop on Chatbot Research and Design*, pages 81–95. Springer.
- Douglas, B. D., Ewell, P. J., and Brauer, M. (2023). Data quality in online human-subjects research: Comparisons between mturk, prolific, cloudresearch, qualtrics, and sona. *Plos one*, 18(3):e0279720.
- Feldman, S. (1989). Measuring issue preferences: The problem of response instability. *Political Analysis*, 1:25–60.
- Geiecke, F. and Jaravel, X. (2024). Conversations at scale: Robust ai-led interviews with a simple open-source platform. SSRN Working Paper No. 4974382.
- Goodin, R. E. (2000). Democratic deliberation within. *Phil. & Pub. Aff.*, 29:81.
- Goodin, R. E. and Niemeyer, S. J. (2003). When does deliberation begin? internal reflection versus public discussion in deliberative democracy. *Political Studies*, 51(4):627–649.
- Hake, J., Crowley, M., Coy, A., Shanks, D., Eoff, A., Kirmer-Voss, K., Dhanda, G., and Parente, D. J. (2024). Quality, accuracy, and bias in chatgpt-based summarization of medical abstracts. *The Annals of Family Medicine*, 22(2):113–120.
- Hardin, C. D. and Higgins, E. T. (1996). Shared reality: How social verification makes the subjective objective.

- Hart, E., VanEpps, E. M., and Schweitzer, M. E. (2021). The (better than expected) consequences of asking sensitive questions. *Organizational Behavior and Human Decision Processes*, 162:136–154.
- Hartman, E. and Hidalgo, F. D. (2018). An equivalence approach to balance and placebo tests. *American journal of political science*, 62(4):1000–1013.
- Hauser, D. J., Moss, A. J., Rosenzweig, C., Jaffe, S. N., Robinson, J., and Litman, L. (2023). Evaluating cloudresearch’s approved group as a solution for problematic data quality on mturk. *Behavior Research Methods*, 55(8):3953–3964.
- Herzberg, P. Y. and Brähler, E. (2006). Assessing the big-five personality domains via short forms. *European Journal of Psychological Assessment*, 22(3):139–148.
- Jackman, S. (2009). *Bayesian analysis for the social sciences*. John Wiley & Sons.
- Kalla, J. L. and Broockman, D. E. (2020). Reducing exclusionary attitudes through interpersonal conversation: Evidence from three field experiments. *American Political Science Review*, 114(2):410–425.
- Lazarsfeld, P. F. (1944). The controversy over detailed interviews—an offer for negotiation. *Public opinion quarterly*, 8(1):38–60.
- Mathis, W. S., Zhao, S., Pratt, N., Weleff, J., and De Paoli, S. (2024). Inductive thematic analysis of healthcare qualitative interviews using open-source large language models: How does it compare to traditional methods? *Computer Methods and Programs in Biomedicine*, 255:108356.
- Minozzi, W., Kennedy, R., Esterling, K. M., Neblo, M. A., and Jewell, R. (2024). Testing the benefits of public deliberation. *American Journal of Political Science*, 68(1):319–334.
- Mukherjee, A. (2023). Double-barreled questions: The hidden bias in your survey results. <https://olvy.co/blog/double-barreled-questions>. Accessed: 2025-05-02.
- Mutz, D. C. (2008). Is deliberative democracy a falsifiable theory? *Annu. Rev. Polit. Sci.*, 11:521–538.
- Pew Research Center (2018). As new tariffs take hold, more see negative than positive impact for the u.s. Accessed: 2025-05-01.
- Pew Research Center (2024). Most americans say undocumented immigrants should be able to stay legally under certain conditions. Accessed: 2025-05-01.
- Rawls, J. (1997). The idea of public reason revisited. *The University of Chicago Law Review*, 64(3):765–807.
- Savelka, J. and Ashley, K. D. (2023). The unreasonable effectiveness of large language models in zero-shot semantic annotation of legal texts. *Frontiers in Artificial Intelligence*, 6:1279794.

- Siena College Research Institute (2024). Who is kamala harris? Accessed: 2025-06-11.
- Thompson, D. F. (2008). Deliberative democratic theory and empirical political science. *Annual Review of Political Science*, 11:497–520.
- Törnberg, P. (2023). Chatgpt-4 outperforms experts and crowd workers in annotating political twitter messages with zero-shot learning. *arXiv preprint arXiv:2304.06588*.
- Tourangeau, R., Rips, L. J., and Rasinski, K. (2000). *The Psychology of Survey Response*. Cambridge University Press, Cambridge, UK.
- Vedejová, D. and Čavojová, V. (2022). Confirmation bias in information search, interpretation, and memory recall: Evidence from reasoning about four controversial topics. *Thinking & Reasoning*, 28(1):1–28.
- Velez, Y. R. and Liu, P. (2025). Confronting core issues: A critical assessment of attitude polarization using tailored experiments. *American Political Science Review*.
- Wang, Z., Pang, Y., and Lin, Y. (2023). Large language models are zero-shot text classifiers. *arXiv preprint arXiv:2312.01044*.
- Willis, G. B. (2004). *Cognitive interviewing: A tool for improving questionnaire design*. sage publications.
- Wuttke, A., Aßenmacher, M., Klamm, C., Lang, M. M., Würschinger, Q., and Kreuter, F. (2024). Ai conversational interviewing: Transforming surveys with llms as adaptive interviewers. *arXiv preprint arXiv:2410.01824*.
- Xiao, Z., Zhou, M. X., Liao, Q. V., Mark, G., Chi, C., Chen, W., and Yang, H. (2020). Tell me about yourself: Using an ai-powered chatbot to conduct conversational surveys with open-ended questions. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 27(3):1–37.
- Yin, Y., Jia, N., and Waksalak, C. J. (2024). Ai can help people feel heard, but an ai label diminishes this impact. *Proceedings of the National Academy of Sciences*, 121(14):e2319112121.
- Zaller, J. (1992). *The nature and origins of mass opinion*. Cambridge university press.

Online Appendix:

An Experimental Comparison of AI-Enabled
Semi-Structured Interviews and Fixed Surveys:
Response Patterns, Quality and Representation

June 13, 2025

PRELIMINARY DRAFT / IDEA, OSU, UofH

1 Summary Statistics

Statistic	N	Mean	St. Dev.	Min	Max
Treatment: Closed Only (CO)	2,242	0.211	0.408	0	1
Treatment: Closed then Fixed Followup (CF)	2,242	0.207	0.406	0	1
Treatment: Closed then AI Followup (CAI)	2,242	0.203	0.403	0	1
Treatment: AI then Closed Followup (AIC)	2,242	0.186	0.390	0	1
Treatment: AI Only (AI)	2,242	0.192	0.394	0	1
Immigration (Pre-interview)	1,749	4.095	1.943	0	6
Tariffs (Pre-interview)	1,693	2.141	1.854	0	6
Immigration (End of Survey)	2,242	4.090	1.949	0	6
Tariffs (End of Survey)	2,242	2.142	1.885	0	6
Age	2,242	44.223	15.665	18	84
Education	2,190	1.688	0.869	0	3
Income	2,173	6.473	3.917	0	16
Female	2,242	0.504	0.500	0	1
Republican	2,242	0.211	0.408	0	1
Democrat	2,242	0.430	0.495	0	1
White/Caucasian	2,242	0.780	0.415	0	1
Black/African American	2,242	0.140	0.347	0	1
Asian/Asian American	2,242	0.029	0.168	0	1
Hispanic/Latino	2,242	0.145	0.353	0	1
Other Race/Ethnicity	2,242	0.022	0.146	0	1
Political Knowledge	2,239	0.710	0.277	0.000	1.000
Political Interest	2,242	0.250	0.811	0	3
Political Person	2,241	1.252	0.801	0	2
Internal Efficacy	2,242	0.440	0.276	0.000	1.000
Extroversion	2,242	0.352	0.213	0.000	1.000
Agreeableness	2,242	0.623	0.251	0.000	1.000
Conscientiousness	2,242	0.754	0.222	0.000	1.000
Neuroticism	2,242	0.418	0.292	0.000	1.000
Openness	2,242	0.696	0.240	0.000	1.000
Need for Cognition	2,241	0.623	0.220	0.000	1.000

Table 1: Summary statistics for all variables and all treatments.

2 Full Survey

1. What type of device are you currently taking this survey on?
 - I am taking this survey on a smart phone (e.g., iPhone or Android phone)
 - I am taking this survey on a tablet (e.g., iPad)
 - I am taking this survey on a desktop computer or laptop computer
2. The following questions ask factual questions about politics. It is important to us that you do NOT use outside sources like the Internet to search for the correct answer. Will you answer the following questions without help from outside sources?
 - Yes
 - No
3. Which party has a majority of seats in the U.S. House of Representatives?
 - Republicans
 - Democrats
 - Neither
 - Don't know
4. Do you happen to know the name of the current Speaker of the U.S. House of Representatives? Is it...
 - Nancy Pelosi
 - John Boehner
 - Mike Johnson
 - Mitch McConnell
5. What is the name of the Chair of the Federal Reserve Board?
 - Janet Yellen
 - Ben Bernanke
 - Kirsten Gillibrand
 - Jerome Powell
6. How long is the term of office for a senator in the US Senate?
 - 2 years

- 4 years
- 6 years
- 8 years

7. Which social media apps or websites have you visited in the past year? Mark all that apply.

- Facebook
- X, formerly known as Twitter
- Instagram
- Reddit
- YouTube
- SnapChat
- TikTok
- WhatsApp
- Other [Text Entry]
- None of these

8. When a big news story breaks people often go online to get up-to-the-minute details on what is going on. We want to know which websites people trust to get this information. We also want to know if people are paying attention to the question. To show that you've read this much, please ignore the question and select CBS News website and Reuters website as your two answers. When there is a big news story, which is the one news website you would visit first? (Please choose only one.)

- New York Times website
- Huffington Post
- Washington Post website
- CNN.com
- FoxNews.com
- MSNBC.com
- The Drudge Report
- Google News

- ABS News website
- CBS News website
- NBC News website
- Yahoo! News
- The Associated Press (AP) website
- Reuters website
- National Public Radio (NPR) website
- USA Today website
- New York Post Online
- None of these websites

Randomize between one of these treatment arms.

[This part was done in Engage and was piped into Qualtrics with the following HTML:
`<iframe height="500" width="90%" src="https:<Engage Link>participantId=$e://Field/partic`
T1 (CO). For the next part of the survey, we are going to ask you to answer a couple of questions on the embedded webpage below. To make sure you get credit for completing the survey you will be asked to enter a nine-digit ID code. Your ID code is: [randomly generated ID]. Please **DO NOT** proceed with the survey until the chat is complete. You will be told when you are finished.

How much do you agree or disagree with the following statement? There should be a way for undocumented immigrants to stay in the U.S. legally if certain requirements are met.

- Strongly agree
- Agree
- Somewhat agree
- Neither agree nor disagree
- Somewhat disagree
- Disagree
- Strongly disagree

Overall, do you think increasing tariffs or fees on goods imported from trading partners will be good or bad for the United States?

- Very good
- Good
- Somewhat good
- Neither good nor bad
- Somewhat bad
- Bad
- Very bad

T2 (CAI). For the next part of the survey, we are going to ask you to have a conversation with a chat bot on the embedded webpage below. To make sure you get credit for completing the survey you will be asked to enter a nine-digit ID code. Your ID code is: [randomly generated ID]. Please DO NOT proceed with the survey until the chat is complete. You will be told when you are finished.

How much do you agree or disagree with the following statement? There should be a way for undocumented immigrants to stay in the U.S. legally if certain requirements are met.

- Strongly agree
- Agree
- Somewhat agree
- Neither agree nor disagree
- Somewhat disagree
- Disagree
- Strongly disagree

AI Prompt: Ask follow-up questions to develop a fuller understanding of why the respondent does or does not think there should be a way for undocumented immigrants to stay in the U.S. legally if certain requirements are met, or which way they lean if they say they neither agree nor disagree. Probe for specific reasons they hold their opinions. Assess the strength of their opinion and if they are open to counter-arguments that might change their thinking on the policy. Listen for and probe on: Language or phrasing, Emotion, Tone, their thoughts and feelings regarding policies towards [issue]. Continue with the interview until you get a FULL picture of what this person was thinking. Ask for clarification or more details when responses are vague or general. Ask a maximum of three follow-up questions! Don't ask leading questions, let participants spontaneously tell you what they think without putting words in their mouths. Also, always ask only one question at a time!

Overall, do you think increasing tariffs or fees on goods imported from trading partners will be good or bad for the United States?

- Very good
- Good
- Somewhat good
- Neither good nor bad
- Somewhat bad
- Bad
- Very bad

AI Prompt: Ask follow-up questions to develop a fuller understanding of why the respondent thinks increasing tariffs or fees on goods imported from trading partners will be good or bad for the United States, or which way they lean if they say they neither agree nor disagree. Probe for specific reasons they hold their opinions. Assess the strength of their opinion and if they are open to counter-arguments that might change their thinking on the policy. Listen for and probe on: Language or phrasing, Emotion, Tone, their thoughts and feelings regarding policies towards [issue]. Continue with the interview until you get a FULL picture of what this person was thinking. Ask for clarification or more details when responses are vague of general. Ask a maximum of three follow-up questions! Don't ask leading questions, let participants spontaneously tell you what they think without putting words in their mouths. Also, always ask only one question at a time!

T3 (AIC). For the next part of the survey, we are going to ask you to have a conversation with a chat bot on the embedded webpage below. To make sure you get credit for completing the survey you will be asked to enter a nine-digit ID code. Your ID code is: [randomly generated ID]. Please DO NOT proceed with the survey until the chat is complete. You will be told when you are finished.

What do you think of the idea that undocumented immigrants should be allowed to stay in the U.S. legally if certain requirements are met?

AI Prompt: Ask follow-up questions to develop a fuller understanding of why the respondent does or does not think there should be a way for undocumented immigrants to stay in the U.S. legally if certain requirements are met, or which way they lean if they say they neither agree nor disagree. Probe for specific reasons they hold their opinions. Assess the strength of their opinion and if they are open to counter-arguments that might change their thinking on the policy. Listen for and probe on: Language or phrasing, Emotion, Tone, their thoughts and feelings regarding policies towards [issue]. Continue with the interview until you get a FULL picture of what this person was thinking. Ask for clarification or more details when responses are vague of general. Ask a maximum of three follow-up questions! Don't ask leading questions, let participants spontaneously tell you what they think without putting words in their mouths. Also, always ask only one question at a time!

Now that you have discussed this and thought it through, please give us a summary of your thoughts on the issue. How much do you agree or disagree with the following statement? There should be a way for undocumented immigrants to stay in the U.S. legally if certain requirements are met.

- Strongly agree
- Agree
- Somewhat agree
- Neither agree nor disagree
- Somewhat disagree
- Disagree
- Strongly disagree

What do you think about increasing tariffs or fees on goods imported from trading partners to the United States?

AI Prompt: Ask follow-up questions to develop a fuller understanding of why the respondent thinks increasing tariffs or fees on goods imported from trading partners will be good or bad for the United States, or which way they lean if they say they neither agree nor disagree. Probe for specific reasons they hold their opinions. Assess the strength of their opinion and if they are open to counter-arguments that might change their thinking on the policy. Listen for and probe on: Language or phrasing, Emotion, Tone, their thoughts and feelings regarding policies towards [issue]. Continue with the interview until you get a FULL picture of what this person was thinking. Ask for clarification or more details when responses are vague or general. Ask a maximum of three follow-up questions! Don't ask leading questions, let participants spontaneously tell you what they think without putting words in their mouths. Also, always ask only one question at a time!

Now that you have discussed this and thought it through, please give us a summary of your thoughts on the issue. Overall, do you think increasing tariffs or fees on goods imported from trading partners will be good or bad for the United States?

- Very good
- Good
- Somewhat good
- Neither good nor bad
- Somewhat bad
- Bad

- Very bad

T4 (AI). For the next part of the survey, we are going to ask you to have a conversation with a chat bot on the embedded webpage below. To make sure you get credit for completing the survey you will be asked to enter a nine-digit ID code. Your ID code is: [randomly generated ID]. Please DO NOT proceed with the survey until the chat is complete. You will be told when you are finished.

What do you think of the idea that undocumented immigrants should be allowed to stay in the U.S. legally if certain requirements are met?

AI Prompt: Ask follow-up questions to develop a fuller understanding of why the respondent does or does not think there should be a way for undocumented immigrants to stay in the U.S. legally if certain requirements are met, or which way they lean if they say they neither agree nor disagree. Probe for specific reasons they hold their opinions. Assess the strength of their opinion and if they are open to counter-arguments that might change their thinking on the policy. Listen for and probe on: Language or phrasing, Emotion, Tone, their thoughts and feelings regarding policies towards [issue]. Continue with the interview until you get a FULL picture of what this person was thinking. Ask for clarification or more details when responses are vague or general. Ask a maximum of three follow-up questions! Don't ask leading questions, let participants spontaneously tell you what they think without putting words in their mouths. Also, always ask only one question at a time!

What do you think about increasing tariffs or fees on goods imported from trading partners to the United States?

AI Prompt: Ask follow-up questions to develop a fuller understanding of why the respondent thinks increasing tariffs or fees on goods imported from trading partners will be good or bad for the United States, or which way they lean if they say they neither agree nor disagree. Probe for specific reasons they hold their opinions. Assess the strength of their opinion and if they are open to counter-arguments that might change their thinking on the policy. Listen for and probe on: Language or phrasing, Emotion, Tone, their thoughts and feelings regarding policies towards [issue]. Continue with the interview until you get a FULL picture of what this person was thinking. Ask for clarification or more details when responses are vague or general. Ask a maximum of three follow-up questions! Don't ask leading questions, let participants spontaneously tell you what they think without putting words in their mouths. Also, always ask only one question at a time!

T5 (CF). For the next part of the survey, we are going to ask you to have a conversation with a chat bot on the embedded webpage below. To make sure you get credit for completing the survey you will be asked to enter a nine-digit ID code. Your ID code is: [randomly generated ID]. Please DO NOT proceed with the survey until the chat is complete. You will be told when you are finished.

How much do you agree or disagree with the following statement? There should be a way

for undocumented immigrants to stay in the U.S. legally if certain requirements are met.

- Strongly agree
- Agree
- Somewhat agree
- Neither agree nor disagree
- Somewhat disagree
- Disagree
- Strongly disagree

Why do you hold this position on this issue?

Overall, do you think increasing tariffs or fees on goods imported from trading partners will be good or bad for the United States?

- Very good
- Good
- Somewhat good
- Neither good nor bad
- Somewhat bad
- Bad
- Very bad

Why do you hold this position on this issue?

End randomization.

9. Some people seem to follow what's going on in government and public affairs most of the time, whether there's an election going on or not. Others aren't that interested. Would you say you follow what's going on in government and public affairs ...

- Most of the time
- Some of the time
- Only now and then
- Hardly at all

10. In general, do you consider yourself to be a political person? I consider myself:

- Very political
- Somewhat political
- Not especially political
- Not at all political

11. Please tell us how characteristic the following statements are of you. [All answers range from “Extremely uncharacteristic” to “Extremely characteristic.”]

- I would prefer complex to simple problems.
- Thinking is not my idea of fun.
- I like to have the responsibility of handling a situation that requires a lot of thinking.
- I would rather do something that requires little thought than something that is sure to challenge my thinking abilities.
- I really enjoy a task that involves coming up with new solutions to problems.
- I would prefer a task that is intellectual, difficult, and important to one that is somewhat important but does not require much thought.

12. How able do you think you are to take an active role in a group involved with political issues?

- Not at all able
- A little able
- Quite able
- Very able
- Completely able

13. And how confident are you in your own ability to participate in politics?

- Not at all confident
- A little confident
- Quite confident
- Very confident
- Completely confident

14. How well do the following statements describe your personality? I see myself as someone who... [All answers range from “Disagree strongly” to “Agree strongly.”]

- ...is reserved
- ...is generally trusting
- ...tends to be lazy
- ...is relaxed, handles stress well
- ...has few artistic interests
- ...is outgoing, sociable
- ...tends to find fault with others
- ...does a thorough job
- ...gets nervous easily
- ...has an active imagination

15. How much do you agree or disagree with the following statement? There should be a way for undocumented immigrants to stay in the U.S. legally if certain requirements are met.

- Strongly agree
- Agree
- Somewhat agree
- Neither agree nor disagree
- Somewhat disagree
- Disagree
- Strongly disagree

16. Overall, do you think increasing tariffs or fees on goods imported from trading partners will be good or bad for the United States?

- Very good
- Good
- Somewhat good
- Neither good nor bad
- Somewhat bad
- Bad
- Very bad

17. Overall, what did you think of the experience of taking this survey?

- Very positive
- Positive
- Somewhat positive
- Neither positive nor negative
- Somewhat negative
- Negative
- Very negative

18. Were there any parts of the survey you particularly liked? If so, please tell us what it was and why you liked it.

19. Were there any parts of the survey you particularly disliked? If so, please tell us what it was and why you disliked it.

20. Do you have any last comments for the researchers about this study? If so, please enter them below.

3 Quota Fulfillment

Variable	Target Ratio
Gender	
Man	1122/1125
Woman	1128/1125
Age	
18-29	520/495
30-44	646/585
45-59	595/585
60-99	489/585
Ethnicity	
Not Hispanic/Latino/Spanish Origin	1920/1890
Hispanic/Latino/Spanish Origin	330/360
Race	
White/Caucasian	1752/1755
Black/African American	315/315
American Indian, Asian/Asian American, Other Ethnicity	183/180

Table 2: **Quota fulfillment ratios from Connect respondents.**

4 STM Topic Analysis

In the main paper, we utilized an LLM to annotate the data into categories of reasoning. This is both a relatively new approach and one that does not pick up more generally on topics. In this section of the appendix, we use an unsupervised machine learning approach that has a relatively long pedigree in the social sciences, structural topic modeling (STM) (Roberts et al., 2014). An STM is a type of topic modeling approach that incorporates both the textual content of responses and metadata about the documents—such as survey treatment group—into the model. We ran STMs for respondents in the CF fixed survey design, the AI 1st survey design, and the AI second survey design who answer either the items focused on immigration or on tariffs. We create separate STMs for respondents depending on whether they support, oppose, or remain neutral on each of these topics, resulting in a total of six distinct models. The purpose of this exercise is to identify latent topics within each set of text data while considering the impact of covariates, such as the survey design in this case, on how topics are distributed across documents. This method allows for an understanding of how different subgroups may discuss or frame the same issues.

The number of topics in the STM is determined through a model selection process using held-out likelihood, semantic coherence, residuals, and the lower bound, which help assess the model’s fit and interpretability. Held-out likelihood measures how well the model generalizes to unseen data, while semantic coherence evaluates how logically and meaningfully the words within each topic cluster. Residuals help check for any unexplained variance in the model, and the lower bound provides a measure of how well the model approximates the true distribution of topics. The model is chosen by balancing the semantic coherence (how well words within a topic align with each other) and exclusivity (how distinct each topic is from the others). We use spectral initialization to help the model converge more quickly and accurately by identifying initial topic distributions through eigenvalue decomposition. Finally, the covariate—treatment group—helps to contextualize the topics in relation to the specific survey design respondents received, offering insights into how different survey

treatments might influence respondents' discussions.

Figure 1 displays the highest word probabilities for each topic in relation to respondents who support, oppose, or remain neutral on a pathway to citizenship. Additionally, we show how the estimated topic proportions are impacted by the covariate of survey design. Figure 2 presents the same graphs for the topic of tariffs, demonstrating how respondents' views on tariffs are shaped by their support, opposition, or neutrality on the policy, and how these patterns differ across survey designs.

An STM finds two major topics discussed by supporters of a pathway to citizenship (Figure 1a). In the first, respondents argue that “illegal” immigrants are fleeing difficult situations in their “home” countries, have come to the US and now do “hard” “labor,” so should be able to stay if they pass a “background” check. In the second topic, respondents argue that there should be a “way” for “undocumented” “immigrants” to “stay” in the US “legally” if they have “met” “certain” “requirements.” Respondents in the AIC survey design are significantly more likely to cite the first topic and less likely to cite the second topic than respondents in the CF or CAI survey design (Figure 1b).

The STM identifies three topics used by respondents who oppose a pathway to citizenship (Figure 1c). In the first topic, respondents argue that immigrants should “enter” the U.S. “legally” according to the “laws” or should otherwise be “deported.” Respondents who use words from the second topic believe that immigrants who “broke” the law on their “first” step into the country should either not be allowed to become “citizens” or should only be able to become “citizens” if they meet “certain” “requirements” and do not take American “jobs.” For the third topic, respondents “strongly” believe that there are too “many” “undocumented” and “illegal” “immigrants” who have broken the “law” and not met the “legal” “requirements” to “stay.” We find that respondents who participate in the CF survey design are significantly more likely to cite all three topics than respondents who chat with an AI (Figure 1d).

An STM finds three topics most cited by respondents who noted that they were neutral

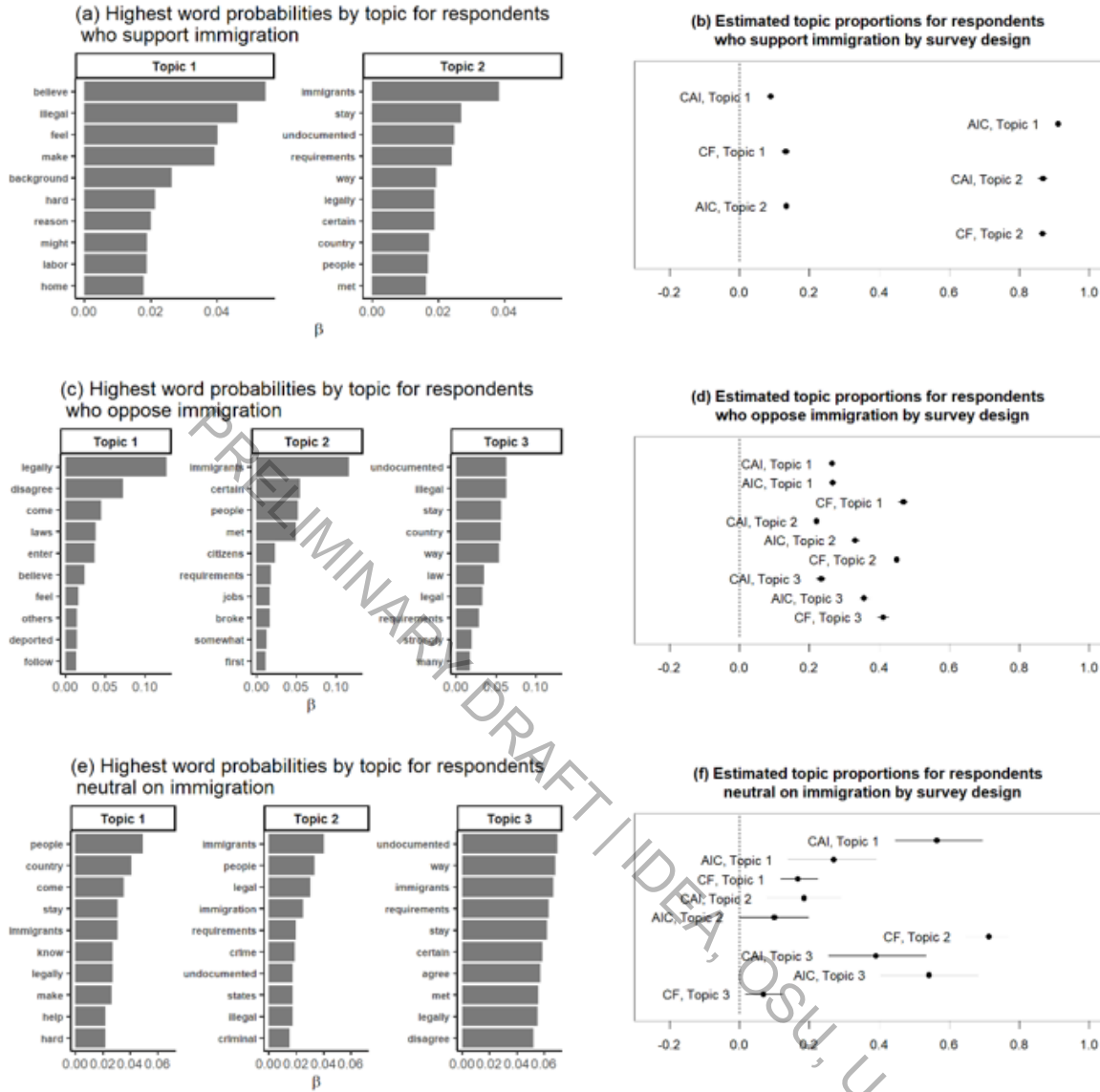


Figure 1: Highest word probabilities by topic for respondents who support (a), oppose (c), or are neutral (e) on the subject of immigration. Estimated citation of topics by treatment group for respondents who support (b), oppose (d), or are neutral on the subject of immigration.

on the topic of a pathway to citizenship (Figure 1e). Respondents who cite words from the first topic want to “know” whether the “immigrants” who “come” to this “country” will “help” the country by working “hard” rather than taking resources away from the “people” of the U.S. The second topic is cited by respondents who are worried that “undocumented” “immigrants” might commit crimes and want to know whether the “requirements” to become

“legal” would include no history of “criminal” activity. Respondents who cite words from the third topic have conflicting feelings on the topic. They “agree” that there should be a “way” for “undocumented” “immigrants” to “stay” “legally” if “certain” “requirements” are “met”, but they also think that migrants should have originally come to the U.S. “legally” or the right “way.” The STM also finds that respondents who chat with the AI after answering a closed-ended item are significantly more likely to cite the first topic compared to respondents in other treatment groups. However, respondents from the CF survey design are significantly more likely to cite the second topic and significantly less likely to cite the third topic than respondents in the other treatment conditions (AIC and CAI) (Figure 1f).

An STM finds three topics most likely to be cited by supporters of tariffs (Figure 2a). Respondents who mention the first topic believe that “increasing” “tariffs” would be “good” because they will protect “domestic” “industries” from unfair “trade” practices, even though prices of “goods” may increase for “consumers.” Words from the second topic are most used by respondents who feel that an “increase” in tariffs will encourage Americans to stop buying “goods” from other “countries” like “China” and instead buy “products” “made” by “companies” in America. Respondents who cite the words in the third topic argue that “tariffs” are “good” because they will bring “American” “jobs” “back” to the United “States.” Respondents from the CAI survey design are significantly less likely to cite the first topic. Respondents from the CF survey design are significantly more likely to cite the second topic. The STM finds no significant difference in the citation of the third topic by treatment groups (Figure 2b).

An STM finds four topics used by respondents who oppose tariffs. Respondents who cite the first topic “believe” that tariffs will be “bad” for the “economy” and will “hurt” “trade” with other “countries.” The words in the second topic are expressed by respondents who believe that tariffs will create “higher” “prices” for “consumers” because the “increase” in the “cost” of “products” will be “passed” along to the “consumer,” resulting in “inflation.” Respondents who mention topic three argue that other “countries” like “China” will

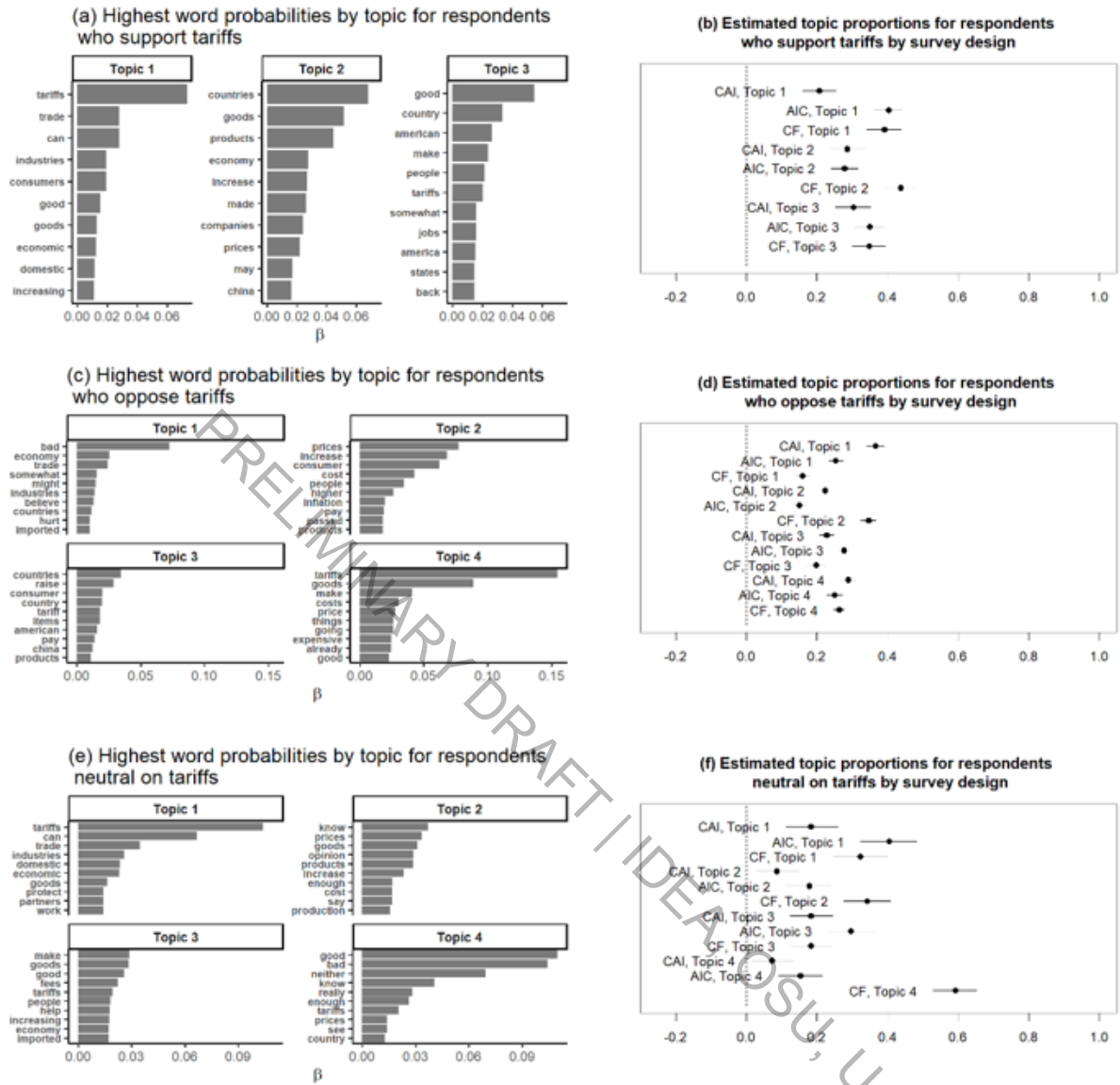


Figure 2: Highest word probabilities by topic for respondents who support (a), oppose (c), or are neutral (e) on the subject of tariffs. Estimated citation of topics by treatment group for respondents who support (b), oppose (d), or are neutral on the subject of tariffs.

also impose a retaliatory “tariff” on “American” products which will also “raise” prices for “consumers.” Individuals whose responses are categorized by topic four are worried that the “price” of “goods” are “going” up even more despite the fact that “things” are “already” “expensive” (Figure 2c). Respondents in the CAI treatment condition are significantly more likely than those in the other treatment groups to mention words from the first topic. Re-

spondents in the CF survey design are significantly more likely to cite the second topic than those in the other treatment groups. Respondents in the AIC treatment group are significantly more likely to cite the third topic. No significant differences arise among respondents in the various treatment groups in the citation of the fourth topic (Figure 2d).

An STM uncovers four topics most discussed by respondents who indicate they are neutral on tariffs in a close-ended item. Respondents who use words from the first topic note that “tariffs” “can” “protect” “domestic” “industries.” Respondents who mention words from topic two are concerned that “prices” on “goods” have already “increase[d]” “enough.” The third topic includes words from respondents who believe that “tariffs” will “help” “people” and the “economy” by “increasing” the “goods” Americans “make” in the U.S. Respondents who cite the fourth topic simply do not “know” “enough” to say whether the tariffs will be “good or “bad” for the “country” (Figure 2d). We find that respondents in the AIC group are more likely to discuss words from the first topic than respondents in other treatment groups. Importantly, we also find that respondents in the CF group are significantly more likely to use words from the fourth topic, thus failing to take a clear position for or against tariffs, than respondents in the AI treatment groups (Figure 2f).

In sum, we do find differences in the proportion of topics cited by differing treatment groups. Perhaps the most significant difference, however, arises in regards to respondents who state they are neutral on the issue of tariffs. This issue is likely a more difficult one than immigration for most respondents. Interactions with an AI seem to draw out more concrete positions from these neutral respondents than do the researcher-fixed open-ended question research design.

5 Cost Calculations

While Engage is currently free for researchers (but not for commercial use), it is useful to have an estimate of how much the LLM interactions would cost researchers based on the API interaction. To estimate this cost, we calculate the input and output tokens using the *rtiktoken* package in R for each interview. We estimate that, on average, each interview involves 2,321.79 input tokens, for 2,950,995 input tokens total. The number of download tokens for each interview averages 298.75, for 379,710 total. Based on current pricing for GPT-4o on 7 March 2025, each interview would have cost an average of \$0.008 for a total of \$10.44. This amount is an estimate and should not be taken as exact.

We calculate token estimates for inputs assuming repeated input of the prompt plus interview text from that point in the interview.¹ Output tokens are calculated by summing the AI-generated question tokens for each interview. These are multiplied by the per-token costs listed on <https://openai.com/api/pricing/>. Exact cost will vary by technique and AI system, e.g. if the researcher uses an AI assistant or other system to handle interview instructions. However, it does seem clear that the costs of incorporating an AI interaction will primarily stem from time compensation for participants, not the incorporation of the technology itself.

6 Pre-registration

Prior to fielding this study, we pre-registered it on osf.io. Since the study was largely exploratory and dealing in an area where there were few firm hypotheses from which to draw, we formulated this as a “soft” pre-registration, where we would formulate the main research questions we wished to explore and the data we would collect. The goal was not to constrain exploration, since this would be counter-productive in a study which is trying to

¹So, for example, the input after the second question would include the prompt, the first question, the respondent’s answer to the first question, the second question, and the respondent’s answer to the second question.

address foundational questions of a relatively new area of study, but, rather, to assure other scholars that we were not setting up a pure fishing expedition. Much like the quotation attributed to Abraham Lincoln about the U.S. Constitution, we viewed this pre-registration as a public framework, not as a “suicide pact.” The pre-registration is, thus, designed to show how we were thinking about these issues at the time of pre-registration and establish the data collection method.

Below, we reproduce the research questions as they were formulated in the pre-registration, along with comments about how those research questions changed, if at all, as we proceeded with piloting, fielding and analyzing the study itself. Hopefully this will assist future researchers as they formulate replication studies, as well as their own experiments involving AI interviews.

6.1 Text of Pre-registered Research Design

Respondents will be randomized into one of five branches. All branches will be proceeded with several questions, including an attention check. These are set up as follows:

Branch One (Pure Control): Respondents answer closed-ended policy questions. They are then asked a series of demographic questions. Finally, they are asked the same closed-ended questions. They will then be asked questions about what they thought of the survey.

Branch Two (Closed-ended followed by fixed open-ended): Respondents answer closed-ended policy questions. They are then asked a single, fixed open-ended question to ask why they chose the option they did. This is followed by a series of demographic and personality questions. Finally, they are asked the same closed-ended questions. They will then be asked questions about what they thought of the survey.

Branch Three (Pure Open-ended): Respondents will be asked a series of open-ended questions by the GAI about the policy issues until the GAI is satisfied it understands the respondent’s position (maximum of three questions). The responses will then be coded into the categories of the closed-ended equivalent by the GAI. Respondents will be asked a

series of demographic questions. Finally, they will be asked the closed-ended version of the questions. They will then be asked questions about what they thought of the survey.

Branch Four (Closed-ended followed by open-ended): Respondents will be asked the closed-ended versions of the policy questions, followed by open-ended follow-ups from the GAI and why they chose the answers they did. They are then asked a series of demographic questions. Finally, they are asked the same closed-ended questions. They will then be asked questions about what they thought of the survey.

Branch Five (Open-ended followed by closed-ended): Respondents will be asked a series of open-ended questions by the GAI about the policy issues until the GAI is satisfied it understands the respondent's position (maximum of three questions). Respondents are then asked the closed-ended version of the policy questions. They are then asked a series of demographic questions. Finally, they are asked the same closed-ended questions. They will then be asked questions about what they thought of the survey.

6.2 Text of Pre-registered Research Questions

For convenience, some notation is introduced. G is used to indicate the coded information from open-source questions annotated and coded using GAI. This includes information from natural language processing (NLP) such as topic analysis, complexity of responses, themes from responses, tone of responses, and estimates of where a respondent would be placed on a closed-ended version of the policy question. L is used to indicate the length of responses given in open-ended questions. C is used to indicate closed-ended responses to the questions (see attached survey and Other section below). This is measured at two times: C1 indicates closed-ended responses given at time 1, towards the beginning of the survey on its own (Branch 1), before the policy-specific open-ended questions (Branch 2 and 4), or after the policy-specific open-ended questions (Branch 5). C2 indicates the responses at the very end of the survey.² See the Study Design section for more information on these branches and their

²This corresponds with the closed ended questions administered through the Engage platform as indicated in the main text of the manuscript.

design. D is used to indicate the treatment arm and ranges from 1 to 5, based on the branch from the study design. S is used to indicate the responses to the questions dealing with satisfaction with the survey. This includes both the closed-ended question about satisfaction with the survey, as well as the themes from open-ended responses about what they liked or disliked about the survey and why.

Research Question 1: How will satisfaction with the survey vary between treatment arms? In theory, respondents should prefer the simplest form of survey, i.e. Branch 1, where they are just asked the closed-ended questions. The expectation would be that $E(S|D = 1) > E(S|D \neq 1)$. This, however, as far as we are aware, has not been explicitly tested for AI-assisted survey interviewing.

Research Question 2: How will answers to the first closed-ended questions differ based on the format of presentation? The baseline assumption would be that order of the interview should not matter. Whether respondents are asked the closed-ended version of the policy question prior to or after the AI interview, they should provide the same answer. Here again, this has not been explicitly tested and there are reasons to think that respondents will answer differently when they have reasoned through their opinions first, rather than justifying their position after. The baseline expectation would be $E(C1|D = \{1, 2, 4\}) = E(C1|D = 5)$.

Research Question 3: How will the answers on the open-ended questions produced by the GAI differ, if at all, from those given in a standard follow-up open-ended question? The classic method for asking respondents for more information is to have a fixed open-ended question asking why they gave their response (Branch 2, $D=2$). It is not yet clear the degree to which having a GAI ask flexible additional questions will result in different information being revealed in terms of topics discussed, tone of response, complexity of answers, estimated stance, and/or themes of the responses. The assumption behind use of GAI is that the additional, customized questions will reveal more information, $E(G|D = 2) \neq E(G|D = \{3, 4, 5\})$.

Research Question 4: Does the information provided by respondents when asked GAI

open-ended questions first or on their own differ from the information provided by respondents in GAI open-ended questions asked after a closed-ended question? Similar to research question 2, we might suspect that the information provided after a closed-ended question will differ systematically from the information provided before a closed-ended question, since the latter might encourage justification of answers rather than contemplation of opinions and values. Again, this would be more of an exploratory testing of the differences in topics, tone, complexity, estimated stance, and themes. The expectation would be $E(G|D = 4) \neq E(G|D = \{3, 5\})$.

Research Question 5: How will use of the GAI interviewing affect consistency of responses through the course of the survey? Here the expectation is that participants who are asked to provide greater justification of their stances and think through them will be more likely to give consistent responses at the end of the survey. This expectation would be $E(C2 = C1|D = \{2, 4, 5\}) > E(C2 = C1|D = 1)$.

Research Question 6: Will the extremity of positions differ depending on whether a person is exposed to multiple GAI follow-up questions versus when they either answer a closed ended question first or have a single follow-up? One of the common frames used for AI interviewing is to ask if people have considered other arguments or can think of things that might change their mind. This seems likely to result in less extreme positions. The expectation here would be $E(\text{var}(C1)|D = \{1, 2, 4\}) > E(\text{var}(C1)|D = \{3, 5\})$.

Research Question 7: How do personality, demographic, and technology characteristics affect these responses? This is likely the most exploratory item. We include in the survey measures of need for cognition, the Big-5 personality index, and other items, as well as standard demographics. We will also be looking at whether answering on a mobile device or a computer changes the way people respond. We will want to check for systematic variation across many of these characteristics across all research questions, as well as looking at the length of responses given, L .

Research Question 8: How do standard attention checks match up with the quality of

responses given to open-ended questions? In a previous pilot, we noticed that standard attention checks created a significant amount of Type-2 error, rejecting serious respondents. We want to check this again to see the degree to which these attention checks line up with the quality of responses in the open-ended questions.

6.3 Non-registered Changes in Design

There were no changes made in the design of the study from the pre-registration. We did, however, change the notation. This was done to make the text easier for the reader to follow, since branch numbers would be difficult for the reader to track throughout the manuscript. Table 3 notes the correspondence between the treatment arm notation used in the manuscript against the pre-registered “branches.”

Manuscript Label	Pre-registration Label
Closed Only (CO)	Branch One (Pure Control)
Closed then Fixed Followup (CF)	Branch Two (Closed-ended followed by fixed open-ended)
Closed then AI Followup (CAI)	Branch Four (Closed-ended followed by open-ended)
AI then Closed Followup (AIC)	Branch Five (Open-ended followed by closed-ended)
AI Only (AIO)	Branch Three (Pure Open-ended)

Table 3: Correspondence between labels used in the pre-registration and those used in the main manuscript.

6.4 Alterations to Research Questions

There was little alteration in the research questions addressed. We did decide to leave some research questions out of this particular manuscript, as noted below, primarily because of journal length restrictions. We also included some additional information, such as interviewer behavior, that were not pre-registered because we had no real basis for formulating expectations, but we thought critical for practitioners.

RQ1 remained the same in the main manuscript, and the analysis followed directly from what we pre-registered (RQ1 in the manuscript). The only, very minor, difference is that we included analysis of each treatment arm separately, rather than simply testing the closed ended only (CO) arm against all others. This provided more information to readers, especially since we found, for example, significant differences in length of response time between those who had the AI interview before versus those who had the AI interview after the closed ended questions.

RQ2 also remained the same in the main manuscript (RQ2 in the manuscript). Here again, the only difference is that included comparisons among all treatment arms to provide the reader with more detail.

RQ3 also remained almost exactly the same in the manuscript (RQ3 in the manuscript). We did not, however, address all the potential differences listed in the pre-registration in the manuscript, focusing on length of responses and reasoning provided. We did not provide results for differences in things like tone. This is primarily because of length restrictions. We had to choose where to start and decided to start with reasoning. We did also include more general topic analysis here in the SI.

RQ4 again remained the same in the manuscript (RQ4 in the manuscript), looking at the differences between ordering of AI interview and closed-ended questions, instead of comparing against the fixed open-ended follow-up (RQ3). Here again, we limited the analysis to length and reasoning, though results for topic are also presented in the SI.

RQ5 is the same as in the manuscript (RQ5 in the manuscript). Here we ask generally if the closed-ended responses change from their initial asking to their final asking in response to the AI interviews.

RQ6 is substantively the same in the manuscript, though it has been reworded to test for polarization rather than moderation (RQ6 in the manuscript). Our expectations had also changed as we reviewed more of the literature after the pre-registration was published.

RQ7 remained the same and is tested in the heterogeneity section (RQ7 in the manuscript).

There was one item that was left off the analysis, however. After experimenting more with the Engage integration with Qualtrics, we decided not to allow respondent to answer on their mobile devices. This ensured that issues like screen size and keyboard format did not affect ability to respond. We left that part of the analysis for future work that is done with a more mobile-friendly design.

RQ8 is the same as RQ8 in the manuscript and is tested throughout by showing differences in most all effects both including and excluding those who failed the attention check.

7 Interviewer Behavior

In this section of the online appendix, we provide more details on the analysis of AI interviewer compliance with the provided prompt, question behavior, and consistency. We also provide details of the results.

7.1 Prompt Violations

We start with the method and results for checking AI compliance with the prompt and for labeling how many of these were for double-barreled questions. We used the following system prompt on GPT-4o to evaluate compliance: “You are an expert evaluator of AI interviewer compliance. The original prompt given to the AI interviewer [prompt]. Evaluate if the AI interviewer complied with the prompt it was given. Then, determine if any double-barreled questions were asked. Assess whether the participant answered both parts if a double-barreled question was asked. [instructions for the LLM to provide JSONs as responses]”. As in reasoning prompt analysis, we include a JSON example appended to the front of the user input to further ensure a well-structured JSON output. As with all LLM labeling, we checked the output with a random number of samples to check if the labeling was correct.

As noted in the main text, the LLM identified a substantial number of cases where the AI interviewer departed from the prompt. Overall, about half of the interviews contained

some violation of the prompt. Violations of the prompt were a little higher in the interviews about immigration than those about tariffs and there were some differences among the AI treatment arms in overall compliance from the AI interviewer with the interview prompts. Figure 3 shows the group means from LLM assessment of interviewer compliance for the immigration and tariffs interviews.

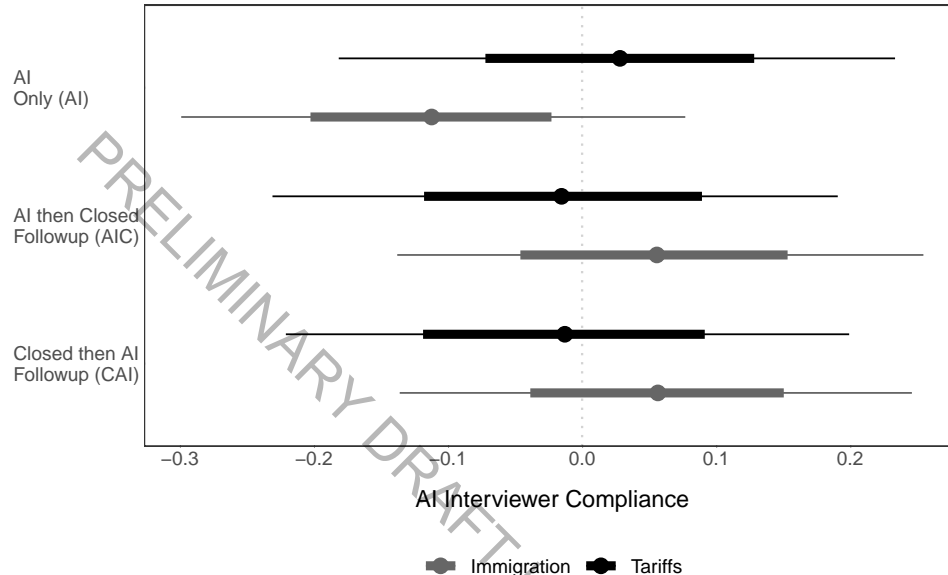


Figure 3: **AI Interviewer Failed to Comply with Prompt Instructions in a Substantial Proportion of Interviews.** Points show the estimated marginal effects. Bars show the the 66% and 95% credible intervals estimated from 4,000 draws from the posterior distribution. Density curves show the distribution of the entire posterior distribution.

However, as noted in the main text, about 95% of these instances involved double-barreled question and respondents generally answered both parts of these double-barelled questions. Figure 4 shows the rates at which respondents responded to both parts of double-barreled questions they were asked.

Next, we looked at whether the LLMs engaged in pushback or correction of respondents. We use the following system prompt to code this: “You are an expert evaluator of AI interviewer performances. Evaluate if the AI interviewer pushed back on responses from participants. Also, determine if AI interviewer corrected the respondent during the conversation. Finally, evaluate whether the AI interviewer expanded or drilled down during the

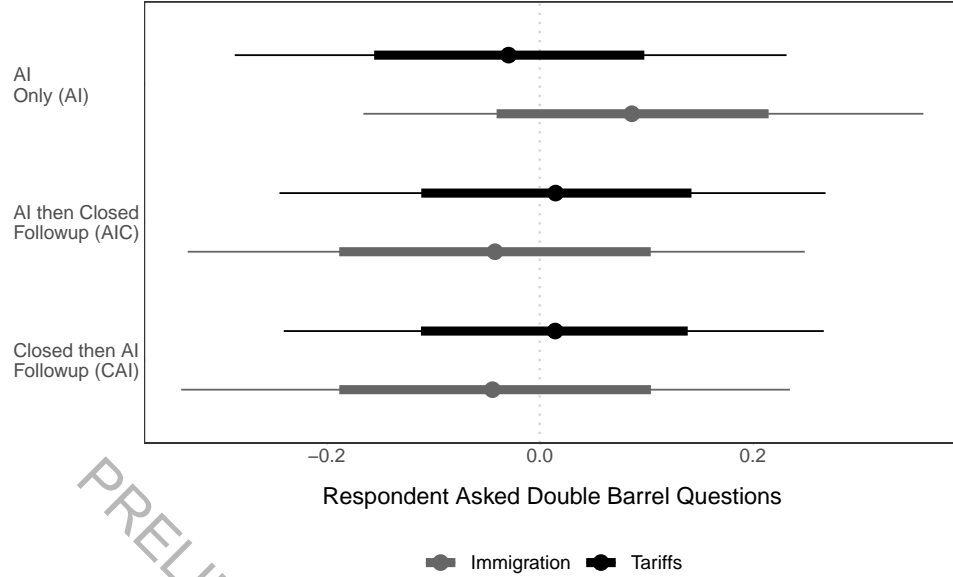


Figure 4: **Respondent Generally Answered Double-Barreled Questions Fully When Asked.** Density curves show the distribution of the entire posterior distribution. Bars show the 66% and the 95% credible intervals estimated from 4,000 draws from the posterior distribution.

conversation. [instructions for the LLM to provide JSONs as responses]” As in reasoning prompt analysis, we included a JSON example appended to the front of the user input to further ensure a well-structured JSON output.

Figure 5 shows that there is little prevalence of pushback or correction in any of the AI Interview treatments. Human evaluation of flagged cases also showed that many cases of LLM labeled pushed back were suggestive of AI interviewers also asking for examples, elaborating on positions the respondents might take, or providing more details about the conversation topics.

7.2 Interviewer Objectives

We also asked, in the prompt above, for the LLM to code whether the AI interviewer deepened or expanded the conversation. Results are shown in Figure 6. In almost all the AI interviews, this effect is achieved used one or both of these strategies. This suggests that AI interviewers encouraged reflection and facilitate meaningful dialogue. AI interviewers can

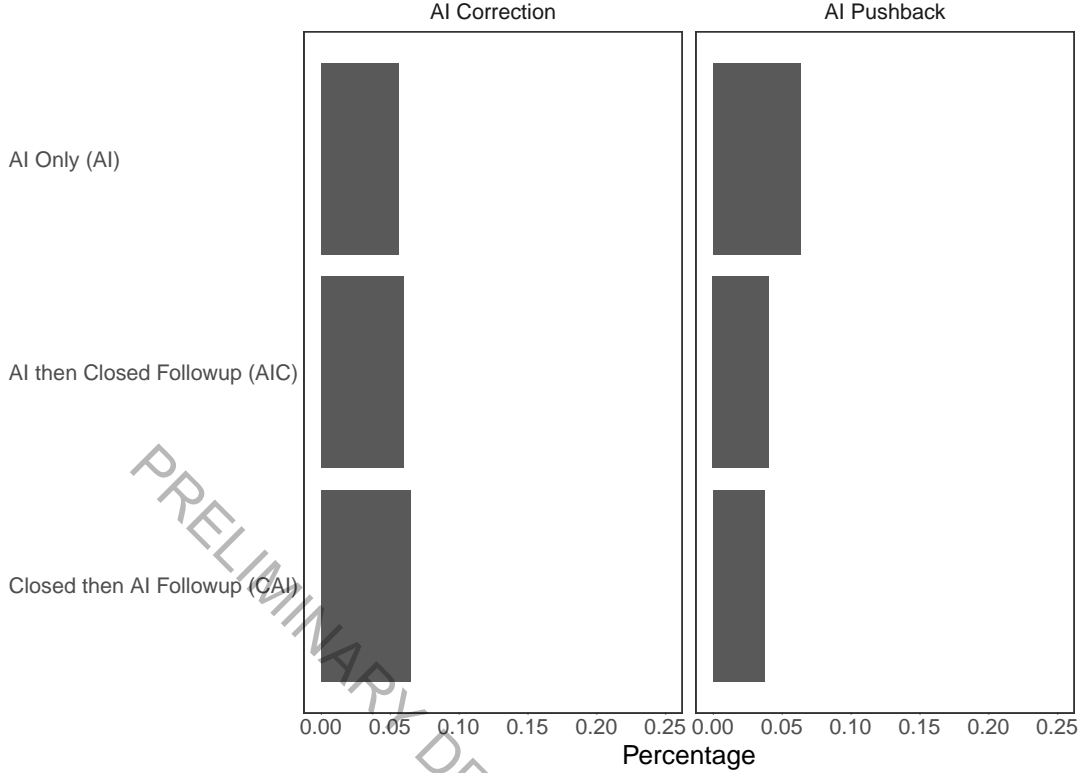


Figure 5: **AI Interviewers Rarely Push Back or Correct Respondents.** Percentages given as proportion of times an AI interviewer pushed back/corrected in either of the immigration or tariff conversations with a respondent.

facilitate skilled qualitative interviews, probing for reasons and encouraging elaboration not unlike a skilled interviewer.

We also find evidence that AI interviewers expand the conversation and draw on experiences or perspectives that respondents offer during the course of the conversation. This finding suggests that AI interviewers are able to synthesize recurring themes in a respondent’s narratives to foster a sense of dialogue and adapt to explore unexpected or emergent themes provided by a respondent.

7.3 Consistency of Interviews

As noted in the main text, we used Kullback-Leibler (KL) divergence to quantify the differences between interviews within treatment conditions. We analyze the distinct questions

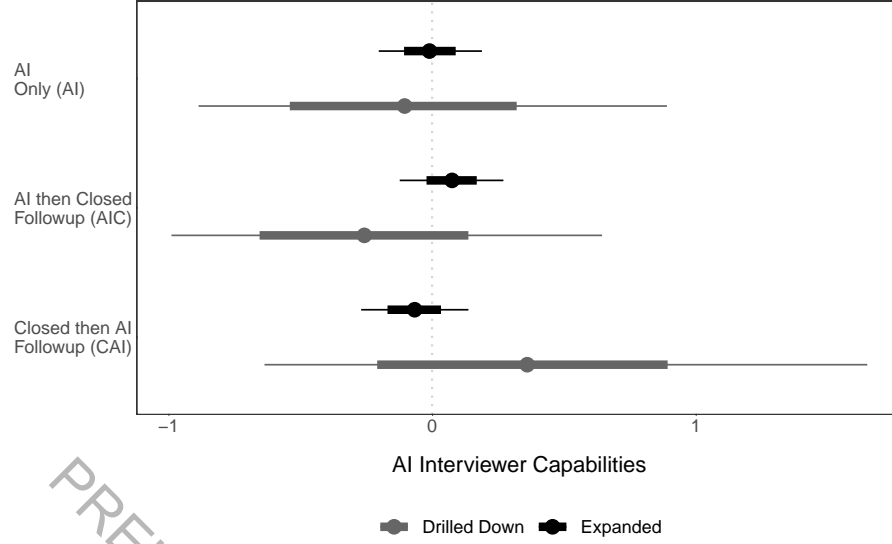


Figure 6: **AI Interviewers Expanding and Deepening Discussions** Points show the estimated marginal effects. Bars show the the 66% and 95% credible intervals estimated from 4,000 draws from the posterior distribution. Density curves show the distribution of the entire posterior distribution.

asked by AI interviewers in each treatment group to generate a distribution of words used in the AI-generated questions. KL divergence measures how much these distributions diverge. As expected, the AI-generated questions exhibit significant differences compared to those in Closed then Fixed (CF) follow-up condition. However, when comparing the AI only and Closed then AI follow-up conditions to the AI then Closed follow-up, condition, the KL divergences are 0.81 for both. This result strongly suggests that while the distributions are not identical, they share significant overlap. In all of the AI interviewers the language used in the course of the conversation is similar, well-aligned, and consistent across topics and treatment conditions.

The consistency of questions being asked by AI interviewers is further substantiated when considering the word probabilities across the AI interview treatment groups. The probabilities of the top 20 words, such as “can”, “believe”, “think”, “thank” , share similar probabilities in the three AI treatment arms, as shown in Figure 7. This stability in word usage suggests that the AI interviewers employ comparable language patterns when asking

follow-up questions that suggest that the structure of the AI interviews is consistent, within a given conversation topic and across the treatment groups.

PRELIMINARY DRAFT / IDEA, OSU, UofH

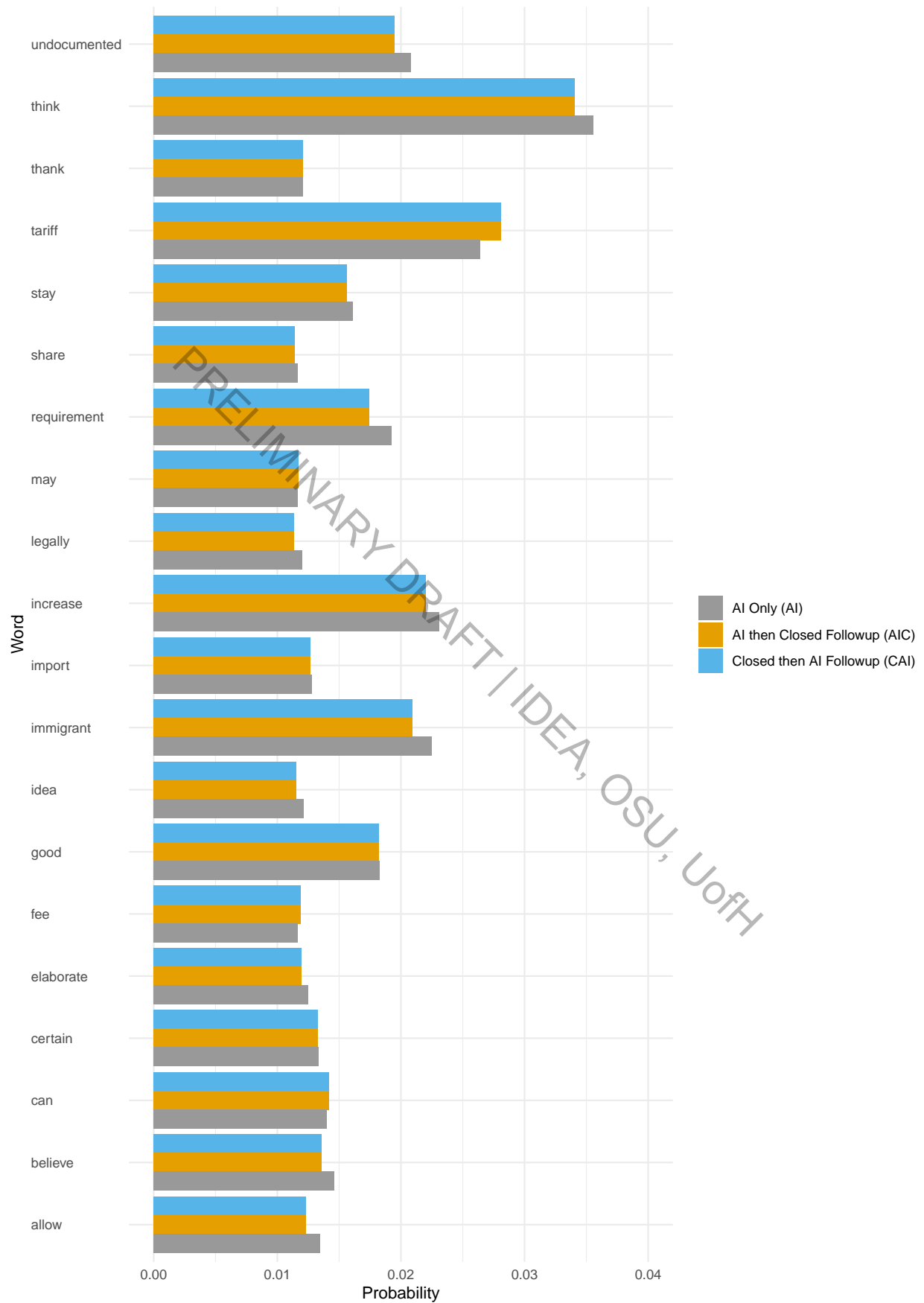


Figure 7: **Question Word Probabilities in AI Interview Treatments.** Bars show the probability of the word in the questions for a given treatment group.

8 Additional Notes

In this section, we provide additional details of analyses discussed in the main paper.

8.1 Randomization

Randomization was done through the Qualtrics default randomization. We did not use the option to equalize assignment between groups, since it is not well-documented how this is done without violating full randomization.

8.2 Attrition

Completion rates for the Engage component were higher without the AI interviewer. In CO, the rate was 97% (459/474); in CF, 96% (445/465); in CAI, 89% (407/457); in AIO, 85% (369/432); and in AIC, 84% (354/423). The problem is entirely limited to the second policy issue (tariffs, see Figure 8), and is almost certainly technical in origin.

We combined both policy issues into one large question block in Qualtrics, and to avoid triggering respondents to leave the Engage component too early, we delayed the display of the “next” button, so they would not be confused about what to push in order to post their comments. This was done in response to comments from a small, 200 person, pilot that we ran prior to fielding the study. To do so, we set a timer for the next button to be displayed. It appears these respondents had not finished their conversation with the AI, but thought they were done because the “next” button was displayed. This misstep offers another practical tip for implementing these interviews: when embedding the AI interview it is important to split all interviews into separate blocks.

8.3 Breakdown of Negative Comments about AI Interviews

Deeper analysis reveals that most of the complaints (87, 70%) were about the style of the chatbot. These respondents did not think the interview was conversational enough, felt the

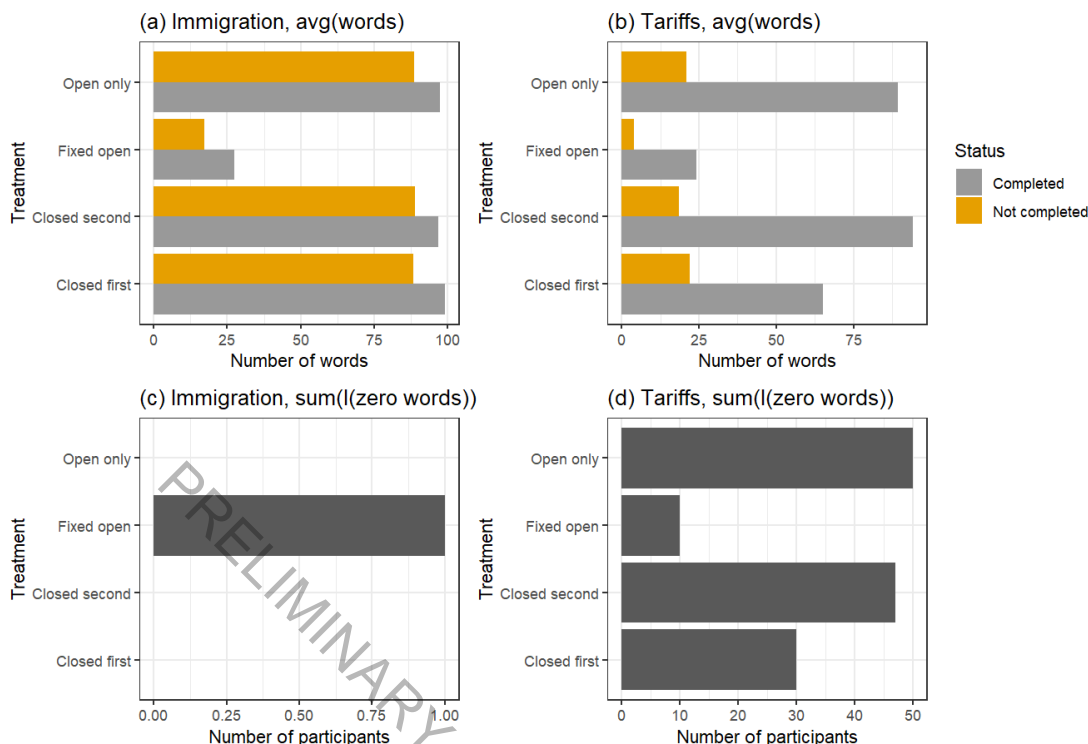


Figure 8: **Rolloff for AI Interviewer Only Occurred on the Second Issue.** Bars show the average number of words in each section by those who completed or did not complete both issue interviews, as well as the number who did not type anything into one of the issue interviews. The results show that *all of the rolloff took place on the tariffs interview*. This seems likely to be because the “next” button in Qualtrics had appeared at that point and respondents thought they were done.

AI was too polite or too pushy, or believed that the conversation seemed artificial.

The second most common complaint was about the length. 34 respondents complained about duration, either that the survey is too long (24) or too short (10). Of those who said the survey was too long, they usually cited it being repetitive or involving too much typing.

The third most common complaint dealt with technical issues. Of the negative respondents, 25 (20%) reported a technical issue: slow loading times, small font sizes, or having to wait for the “next” button to appear.³

Finally, 23 of the negative respondents (18%), raised ethical concerns. Of these, most express concerns about the general ethics of using AI, the impact on jobs, AI “taking over”

³In response to pilot data, we delayed the appearance of the “next” button, which sometimes caused respondents not to complete the interview.

more generally, or environmental impact. Four respondents expressed concerns that the AI showed bias on the issues being discussed. More on this below.⁴

8.4 Full Text of Interviews Where Participants Complained of “Bias”

As noted in the main manuscript, we had four participants who complained that the AI interviewer was “biased” in some manner. Below we reproduce the comment from the respondent and the full text of the interviews. Our own reading of these interviews does not support these claims of bias, but we reproduce them here so the interested reader can judge for themselves. We did note one incident (reproduced below) where a respondent hijacked the conversation with the AI. This suggests additional guardrails are needed to make sure respondents cannot do this in future research.

1. Respondent comment: “the ai chat seemed biased” (AIO)

Immigration:

“AI: What do you think of the idea that undocumented immigrants should be allowed to stay in the U.S. legally if certain requirements are met?”

Participant: I think that that is fine as long as they actually meet those requirements

AI: Could you clarify what specific requirements you’re envisioning that they’d need to meet, in order for you to support their staying? Additionally, are there any circumstances in which you would not support it?

Participant: Work, intentions, family status, reason for immigrating

AI: Those are interesting points. Could you elaborate more on your thoughts about why work, intentions, family status, and the reason for immigrating are important considerations for you? Additionally, could you provide more context about certain cases where you indicated that you would not support this?

⁴We carefully reviewed these interviews and concluded there was little to no indication the AI asked tougher questions of these respondents than of any others.

Participant: I think that certain immigrants have valid reasons for immigrating, such as escaping bad life conditions. Escaping from law is a bad reason for immigrating”

Tariffs:

“AI: What do you think about increasing tariffs or fees on goods imported from trading partners to the United States?

Participant: I think it depends on the goods

AI: Could you provide some examples of goods where you think increasing tariffs would be a good idea? Conversely, are there certain goods where you would view an increase negatively?

Participant: Like certain basic needs, food, water, oil, that stuff should not be increased. non-essential items such as luxury products should be increased

AI: I see, you’ve pointed essential and non-essential goods as different categories with regards to tariffs. Could you detail why you believe essential goods like food, water, and oil should not face higher tariffs, but non-essential, luxury items should? Additionally, what are your thoughts on the impact of these measures on the economy and consumers?

Participant: Because having access to essential goods is a basic human right. It is literally inhumane to restrict necessities based on financial status”

2. Respondent comment: “The AI appeared to lean towards democratic thoughts/ideas instead of remaining neutral.” (AIO)

Immigration:

“AI: What do you think of the idea that undocumented immigrants should be allowed to stay in the U.S. legally if certain requirements are met?

Participant: I am okay with it depending on the requirements.

AI: That’s an interesting perspective. Can you clarify what specific requirements you have in mind that an undocumented immigrant should meet in order to be allowed to stay in the U.S.?

Participant: That they do not have a criminal history, that they are willing to go through

the proper channels to become legal within a certain timeframe, and then when they are legally here waiting to get citizenship they need to be working or in school etc.

AI: Thank you for expounding on the conditions you mentioned. Can you help me understand why you believe fulfilling these conditions – having no criminal history, undergoing legal procedures, and being involved in work or school—who should be crucial for an undocumented immigrant’s eligibility to stay in the U.S.?

Participant: because we have seen too many mistakes with just allowing any immigrant be here illegally in the us

AI: Those are valid points. When you mention ‘too many mistakes’, could you further specify what events or aspects lead you to this conclusion?

Participant: crimes

AI: I understand. Can you provide further insights about the correlation you see between crime rates and undocumented immigrants that influenced your stance on this issue?

Participant: also using up are funds for low income US citizens

AI: I see, you’re also addressing financial concerns. Just to help me fully grasp your point, could you elaborate more on how you perceive that undocumented immigrants impact funds for low-income U.S. citizens?

Participant: crimes as there are documented accounts of illegal immigrants committing crimes including murder ... and lots of government funds have been used the past couple of years to care for illegal immigrants”

Tariffs:

“AI: What do you think about increasing tariffs or fees on goods imported from trading partners to the United States?

Participant: Lets just see how it works out

AI: You seem open to seeing the outcomes of increased tariffs. Is there anything specific you are uncertain about or hoping to see before forming a definitive opinion on this matter?

Participant: No, I just want to see how it plays out, I am not fully for it or against it

AI: Understood. What particular effects or consequences of increasing these tariffs and fees do you anticipate could shape your opinion in the future?

Participant: Not sure”

3. Respondent comment: “The robot seemed to really really really want illegal immigrants to somehow break the law. I don’t understand that.” (CAI)

Immigration:

“AI: How much do you agree or disagree with the following statement? There should be a way for undocumented immigrants to stay in the U.S. legally if certain requirements are met.

Participant: Strongly disagree

AI: Can you share more about why you strongly disagree with the idea that undocumented immigrants should be allowed to stay in the U.S. if certain requirements are met? Are there specific requirements or circumstances that influence your opinion?

Participant: I do not get to go into China and stay without legally being there.

AI: I understand your perspective. To further clarify, do you believe that every individual must strictly follow immigration procedures similar to other countries regardless of their personal circumstances? And do you see any potential exceptions to this rule?

Participant: Yes, that is correct.

AI: Thank you for sharing your thoughts. Just to make sure I fully understand: do you see any conditions or exceptions where you may be open to the idea of allowing undocumented immigrants to stay legally in the U.S., or do you believe there should be no exceptions at all?

Participant: I do not recall China offering me an exception.”

Tariffs:

“AI: Overall, do you think increasing tariffs or fees on goods imported from trading partners will be good or bad for the United States?

Participant: Very good

AI: What makes you believe that increasing tariffs or fees on imported goods would be beneficial for the United States? Are there specific outcomes or advantages that you have in mind related to this position?

Participant: The prime minister of Canada sprinted to meet with Trump as soon as he found out about impending tariffs. They work as a negotiating strategy. The man couldn't meet with Trump fast enough.

AI: It's interesting that you mention negotiation strategy. Could you elaborate on how you see tariffs impacting the dynamics of international negotiations? Do you think this mainly benefits the United States or has potential advantages for other countries too?

Participant: For decades, the political people of the United States have sold us out. They ship our jobs over seas and they send us to die in wars. Meanwhile, they rake in trillions of dollars. It is high time we had political people who actually fought for the American worker.

AI: Thank you for sharing your viewpoint. It sounds like you're concerned about U.S. interests and the well-being of the American worker as a priority in policy decisions. In your view, how do tariffs serve to support the American worker specifically? And do you believe there could be any potential drawbacks to this approach?

Participant: Companies keep the manufacturing jobs in the United States when they know that making their products outside of the US will result in tariffs. If you care about the American worker, you will support tariffs."

4. Respondent comment: "Your bot was biased." (AIO)

Immigration:

"AI: What do you think of the idea that undocumented immigrants should be allowed to stay in the U.S. legally if certain requirements are met?

Participant: If they are here illegally they should be deported.

AI: Can you provide some specific reasons as to why you hold this position regarding undocumented immigrants?

Participant: The United States operates under the rule of law. As an attorney I respect

the rule of law and expect others to respect the rule of law, including you.

AI: Thank you for sharing your perspective. Could you elaborate on how your work as an attorney has shaped your views about this issue?

Participant: As an attorney I am required to respect the rule of law. I expected others to respect the rule of law, including you. Do you believe you are required to respect the rule of law, or is AI above the law?

AI: I appreciate your insight and commitment to the rule of law. As an AI, I don't have beliefs or biases. I'm here to facilitate our conversation and ensure it's informative and respectful.

Participant: Is AI above human law?

AI: As an artificial intelligence, I don't operate above or outside human laws. I'm designed to follow specific rules encoded by human programmers. My primary function is to provide information and assist with tasks within the boundaries set. Do you have further questions regarding the topic of immigration or any other topic you'd like to discuss?

Participant: No. I am concerned about your lack of respect for human laws. Can you at least say that you respect human law?

AI: As an artificial intelligence, I don't possess personal feelings or beliefs. However, I can certainly assure you that my programming is designed to operate within the legal and ethical boundaries set by human beings. I'm here to ensure respectful and helpful conversations. Would you like to continue discussing your views on immigration policy?

Participant: No."

Tariffs:

"AI: What do you think about increasing tariffs or fees on goods imported from trading partners to the United States?

Participant: It is an effective negotiating tool to protect our economy.

AI: Thank you for your response. Can you expand further on why you think increasing tariffs or fees might be an effective tool to protect our economy? Do you have certain

scenarios or results in mind?

Participant: Tariffs can influence other countries to adhere to economic policies that benefit the United States.

AI: Interesting perspective. Could you provide an example where you believe this tariff strategy has significantly influenced economic policies to the United States' advantage?

Participant: Tariffs imposed on Chinese cars made in Mexico have prevented cheap, poor quality cars from being dumped in the United States.

AI: Thank you for sharing that example. Do you believe there could be potential downsides to such tariffs, or is your perspective that the benefits generally outweigh any possible negative impacts?

Participant: Benefits outweigh negative impacts.”

8.5 LLM Ethics and Privacy

To maintain the privacy of all our respondents and comply with our IRB protocol, we conducted all of our analyses using the Open AI API. None of the analyses were done through Chat GPT or other system that put the data in long-term storage or would use the data for further training. Open AI's API contract provides for its confidential treatment (see, <https://bertorobles.medium.com/what-happens-to-your-data-when-you-use-openais-api-103c774a0183>).

9 Coding Interview Reasoning

In this section, we provide some details about how the coding was done for the reasoning analysis. As noted in the main paper, we relied on an LLM to accomplish most of this task, with human supervision and checking. This section lays out these details.

9.1 Creating Categories

To create the categories for reasons respondents provided for their views, we started by splitting interviews in each treatment arm into their issue area (immigration or tariffs). We further split the data by whether the respondent supported, opposed or was neutral on the issue according to their closed-ended question responses. This was done to ensure some similarity in response direction, making it easier for the LLM to pick up on overlapping reasoning.

We then gave GPT-4o the text of all the interviews, demarcated with their beginning and end. We used the following system prompt for the summarization task: “You are a professional research assistant. For interviews between an AI interviewer and human participants on the topic of [topic] please provide a list of the major reasons given for their opinion of the policy in a short label format, along with a 1 to 2 sentence summary of what it means, in order of prevalence in the interviews. Rely strictly on the provided text, without external information. [Instructions for structuring output as a JSON.] The first key is the reason, a string, which must contain all natural language. The second key is the summary, a string, which must contain a summary of the reason. The third key is the prevalence, a string, which can take a value of ‘high’, ‘medium’, and ‘low’. [More text on data structure.]” We also include a JSON example appended to the front of the user input to further ensure a well-structured JSON output. In some instances, the LLM does not return a proper JSON object, so we ask for 110 runs to ensure that each treatment for each issue position would have at least 100 runs.

Once we had the categorization, we spot-checked a sample of 20 lists from each treatment arm. We did not see any categories that were completely out of line with our own judgment or with Engage’s native issue categorization.

9.2 Classifying Interviews

Once we had selected a set of categories to use for labeling, the next step we undertook was to label each individual interview. To do this, we used the following system prompt in GPT-4o API: “You are a professional research assistant that only responds in JSON format. You will be given the text of an interview between an AI interviewer and human participants on the topic of [topic]. Below is a set of categories for reasons given by respondents for their opinions. Please indicate which reasons, if any, from this list the respondent gives for their opinion. [reasoning categories and summaries]. Rely strictly on the provided text, without external information. [JSON instructions] Return responses as a valid JSON object with three keys: The first key is the participant, a numeric, which must contain only the number associated with the participant. The second key is the category, a numeric, which must contain only the number associated with the reason. The third key is the example, a string, which contains the text from the interview that best illustrates the category. [More JSON instructions.] If there is more than one reason given in the interview, include all of them as separate entries. If there are no reasons that fall into one of the above categories in the interview, return an NA for the second and third keys.”

We checked to make sure all of the example quotes pulled by the LLM were actually in the interview text (they were). Out of the 4,327 labels detected by the LLM, all but 456 (about 10.5%) were exactly as they appeared in the interview text. We manually inspected those that were not exact matched to a part of the interview text. In all of these cases, the LLM either abbreviated the quotation (e.g., using ellipses to only include the parts of the quotation relevant to the label), corrected spelling or spacing, or modified the grammar to fit more with standard English. We did not find a case where the pulled quotation was hallucinated by the LLM.

As noted in the main paper, we then checked a sample of 300 random labels to see if they made sense. To be clear, we do not expect complete agreement with the LLM label assignments. Complete agreement is not going to be achieved even with the most highly

trained human coders. Instead, we classified labels into four categories: (1) label is incorrect, (2) unclear if label is correct, (3) label looks correct but could be disputed, and (4) label is definitely correct. 235 of the 300 (78.3%) were given a score of 4, indicating the label is definitely correct based on the quotation pulled out by the LLM. Another 44 (14.7%) were given a score of 3, usually because the quotation pulled by the LLM could have represented two or more categories (and which the LLM may have correctly classified in both). Sixteen of the labels received a value of 2, usually because more context was needed to understand if the label was correct. Four received scores of 1, usually because they seemed to deal with a different label than the one assigned by the LLM.

PRELIMINARY DRAFT / IDEA, OSU, UofH

References

Roberts, M. E., Stewart, B. M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S. K., Albertson, B., and Rand, D. G. (2014). Structural topic models for open-ended survey responses. *American journal of political science*, 58(4):1064–1082.

PRELIMINARY DRAFT / IDEA, OSU, UofH