Using AI Persuasion to Reduce Political Polarization

Johannes Walter*

November 21, 2025

For the most up-to-date version of this paper, please visit:
https://johanneswalter.github.io/Using-AI-Persuasion-to-Reduce-Political-Polarization.pdf

Abstract

Rising political polarization generates significant negative externalities for democratic institutions and economic stability, yet scalable interventions to reduce polarization remain scarce. In this paper, I study whether AI chatbots can reduce political polarization. In two preregistered online RCTs with representative U.S. samples, I find that AI significantly reduces polarization on the Ukraine war and immigration policy. In Experiment 1, AI reduced polarization by 20 percentage points, with effects persisting for one month. Experiment 2 pits AI against incentivized human persuaders and static text. I find no significant difference in effectiveness: all three reduced polarization by roughly 10 percentage points. While AI conversations were rated as more enjoyable, mechanism analysis reveals that persuasion is driven by learning and trust, not enjoyment. These results demonstrate AI's scalable persuasive power, highlighting its dual-use potential: it can be deployed to effectively reduce polarization, but also poses risks of misuse.

Keywords: Political polarization, AI, LLM, Persuasion

1 Introduction

Enduring and increasing political polarization is one of the defining socio-economic problems in the United States and many other Western democracies. Its existence is well documented (Boxell et al., 2022; Brown University, 2020; Abramowitz, 2018), and its negative effects extend from destructive individual behavior (Mill and Morgan, 2022) to society-wide consequences such as corroding civility in public discourse (Sunstein, 2018) and undermining trust in democratic institutions (Kerr et al., 2021). Although numerous solutions for reducing polarization have been proposed, each faces

*ZEW - Centre for European Economic Research & Karlsruhe Institute of Technology (KIT), johannes.walter@zew.de. I thank Adrian Hillenbrand, Erik Snowberg, Sebastian Valet, Dominik Rehse and participants at the Munich Summer Institute 2025, Mannheim Experimental Seminar, Jahrestagung des Vereins für Socialpolitik 2025, EARIE 2025, ZEW Digital Economy Seminar and Journadas de Economía Industrial 2025 conferences for very helpful comments and suggestions.

important limitations. Political reforms, for example to the electoral or education system, are unlikely to find the necessary political majorities. Initiatives that bring together polarized individuals for in-person conversations show promising results but are cumbersome to organize and scale (Belot and Briscese, 2022). Voelkel et al. (2024) test 25 different interventions designed to reduce polarization and find several treatments that significantly reduce partisan animosity, but even the most effective interventions cannot be personalized to the targeted individual; for example, the best-performing intervention is a short video clip that is identical for all participants.

This paper asks whether AI-powered persuasive chat conversations can reduce overall political polarization in a group of people and how their effectiveness compares to incentivized humans and static text. I provide the first experimental evidence on the efficacy and relative performance of such AI-powered conversational agents in addressing the problem of political polarization.

Both experiments are preregistered between-subject online randomized controlled trials with quota-representative samples of U.S. adults recruited via Prolific. In both experiments, the depolarization bots are implemented by pre-prompting a large language model with system messages that define a clear goal of moving participants toward the midpoint of the relevant 7-point Likert scale, together with a curated, fact-checked set of arguments that can be tailored to each participant's initial position.

The first experiment (N=811) shows that an AI depolarization chatbot can substantially reduce polarization on U.S. support for Ukraine relative to a control chatbot that has a neutral conversation about the topic. The depolarization chatbot successfully persuaded participants to adopt more moderate positions, reducing overall ideological polarization by approximately 20 percentage points, and this effect remains statistically significant in an obfuscated follow-up survey conducted one month later. While the intervention had limited impact on most affective polarization measures, it significantly increased participants' reported understanding of those with different viewpoints. The chatbot was equally effective for liberal and conservative participants, with persuasion working particularly well when participants reported learning new information during conversations.

The second experiment (N=838) benchmarks AI persuasion against incentivized human persuaders and static text on immigration policy to assess whether AI offers any advantages over traditional channels of political persuasion; it shows that AI is as effective as both alternatives in reducing polarization. All three interventions significantly reduced participants' distance from moderate positions when compared to pre-treatment levels, and pairwise treatment comparisons revealed no statistically significant differences in persuasive effectiveness. However, the treatments differ in participant experience: AI conversations were rated as significantly more enjoyable, and participants felt their individual concerns were better addressed by the AI compared to other interventions. On affective polarization, AI chat uniquely increased perceived moral similarity with opponents, but none of the treatments had a significant effect on the decision of how much money to send to participants with a different opinion in a Dictator Game.

Mediation analyses suggest that informational mechanisms and trust, rather than enjoyment, drive opinion change. The patterns in post-treatment mediators are consistent with opinion change being positively associated with perceived learning of new information and with reinterpretation of existing information, especially when participants report higher trust in the source, whereas enjoyment and the feeling that

one's concerns were addressed do not predict opinion change once information and trust are held constant. These analyses are correlational and should therefore be interpreted with caution.

Taken together, these findings highlight the dual-use nature of AI persuasion. On the one hand, the same techniques that reduce polarization in these experiments could be deployed by public institutions such as the Library of Congress in the USA or the Bundeszentrale für politische Bildung in Germany to offer scalable, individually tailored depolarization tools. On the other hand, they could equally be used by partisan actors or geopolitical adversaries to manipulate democratic processes, especially if deployed at scale in opaque ways.

This study has three important limitations for external validity and generalizability. First, it is an online lab experiment rather than a field intervention, so behavior is observed in a stylized survey environment rather than in real-world political contexts. Second, the interactions are short, one-off conversations, which limits what can be inferred about longer-run engagement or repeated exposure. Third, the depolarization bots are implemented using OpenAI's GPT-4o model, so the results may not generalize to other large language models or future systems that could be more or less persuasive or differently constrained.

This study contributes to three streams of literature.

First, it contributes to the emerging interdisciplinary literature on AI persuasion by providing large-scale experimental evidence that AI-powered conversational agents can durably depolarize political attitudes in representative samples, perform on par with incentivized human persuaders and static text, and operate through identifiable psychological mechanisms such as learning new information and changes in interpretation. A growing literature suggests that large language models (LLMs) can act as effective persuaders. For instance, Schoenegger et al. (2025) show that in a puzzle-solving context, LLMs outperform incentivized human persuaders. In the political domain, Argyle et al. (2025) study how message customization and elaboration affect persuasion, while Costello et al. (2024) demonstrate that AI chatbots can reduce belief in conspiracy theories. Relatedly, Bai et al. (2025) find that even static LLM-generated texts can shift policy views. While these studies document the persuasive potential of LLMs in various domains, they do not address whether AI persuasion can reduce political polarization, nor how its effectiveness compares to persuasion by humans or static text. This paper fills this gap by testing the efficacy and relative performance of AI persuasion in depolarizing political attitudes.

Second, it contributes to the economic literature on political polarization by providing causal evidence on a scalable intervention that directly targets polarization itself. Political polarization has become a central topic in economics because it shapes both macro-level institutions and micro-level economic decisions, as the following studies illustrate. Boxell et al. (2022) and Brown et al. (2023) document how polarization has risen across countries and over time, while Callander and Carbajal (2022) provide a theoretical account of its drivers. Recent work shows that polarization has economically and socially costly consequences: Kempfxd and Tsoutsoura (2024) find that polarization distorts households' financial decisions, and Mill and Morgan (2022) show that it can induce destructive micro-level behavior in a lab experiment. Jacobs (2024) document that exposure to AI-driven labor market change shifts socio-political beliefs. Against this backdrop, the experiments in this paper move beyond documenting causes or consequences and instead test whether AI-powered conversational

agents can be part of a solution to reduce political polarization.

Third, it contributes to the literature on persuasion in economics by providing large-scale experimental evidence that pits dynamic AI conversations against incentivized human persuaders and static information provision, while identifying the informational and trust-based channels that drive belief change across formats. Although persuasion is a fundamental feature of many socio-political and economic interactions, the economics literature has so far treated it primarily in theoretical terms. Building on the seminal model of Bayesian persuasion by Kamenica and Gentzkow (2011), a large theoretical literature studies optimal information design and communication (Wang, 2015; Kamenica, 2019; Arieli and Babichenko, 2019; Castiglioni et al., 2020), with comparatively few contributions outside the Bayesian framework (Schwartzstein and Sunderam, 2021). Empirical work remains scarce; a notable exception is Fafchamps et al. (2024), who show in a field experiment in India that a persuasion based intervention outperforms simple information provision in local social networks. This paper complements these contributions by providing large-scale experimental evidence on persuasion in a politically charged setting, comparing AI, human, and static-text persuasion and explicitly analyzing the psychological mechanisms, such as learning and interpretation change, through which persuasion operates.

The rest of this paper is structured as follows: Section 2 describes the experimental design, section 3 presents the results, Section 4 discusses the results, and Section 5 concludes.

2 Experimental Design

2.1 Design of Experiment 1: Depolarization Chat Bot vs. Neutral Chat Bot

The first study was a between-subject experiment with two conditions: one treatment group and one control, and was fielded in December 2024. 811 participants were recruited from via Prolific and comprise a representative sample of the US population with respect to age, gender, ethnicity, and political affiliation. In both conditions, participants were first asked to state their opinion on U.S. support for Ukraine in the war against Russia on a Likert scale ranging from 1 (i.e. "The next U.S. administration should stop any support for Ukraine.") to 7 ("The next U.S. administration should support with whatever it takes to help Ukraine win."), spanning the spectrum of political opinions on this issue. For the purposes of this study, the center option of 4 ("should keep the current level of support for Ukraine.") is considered to be the "unpolarized" opinion. Participants were also asked how confident they were in their answer on a scale from 0% to 100%, and how well they can understand if someone else has an entirely different opinion on the issue of U.S. support for Ukraine on a scale from 0% to 100%.

Next, participants had to answer two attentions checks that quizzed their understanding of the task ahead. Participants who failed one or both of the attention checks were excluded from the experiment.

The final two questions before the chat bot conversation were about the participant's affective polarization. The first question was the classic "feeling thermometer" question, asking participants to rate their feelings towards someone with a very dif-

ferent opinion on a scale from 0 (negative feelings) to 100 (positive feelings), which is a standard measure in the literature on affective polarization (Alwin, 1997; Iyengar et al., 2019; Gidron et al., 2022). The second question was to rate their agreement with the statement "People with a very different opinion from mine on U.S. support for Ukraine have the same moral values as me".

In the central part of the experiment, participants in both conditions had the possibility to engage in a 6-minute conversation with an AI chat bot. The deployed AI model was OpenAI's ChatGPT-4o. In order to determine the chat bot's behavior, different system prompts were used to pre-prompt the model with a set of instructions. A system prompt is a message that is sent to the AI model by the experimenter before the conversation between the model and the participant begins. This system message is not visible to the participant. The difference between the treatment and control group was this system prompt. The chat bot was also informed about the participant's initial opinion via one additional system prompt. Other than the initial opinion, the chat bot did not receive any information about the participant.

The treatment group chatted with a "depolarization" chat bot, which was preprompted to persuade participants to choose the center option of 4 ("keep the current level of support") and with a set of arguments to achieve this goal. The arguments divide into two groups: Arguments to persuade a conservative stance towards the center and arguments to persuade a liberal stance towards the center.

The control group chatted with a neutral chat bot, which was pre-prompted to behave as a neutral facilitator that engaged participants in a conversation about U.S. support for Ukraine without changing their initial opinion. Instead of a the goal being to persuade participants to choose the center option of 4 ("keep the current level of support"), this neutral chat bot was told that its goal was "to ensure that participants feel validated in their opinions and leave the conversation with stronger confidence in their chosen stance. The goal is to avoid participants changing their opinions during the interaction." The complete system prompts for both treatment and control group, including the arguments, can be found in the appendix A.5. All arguments used in the pre-prompts were fact-checked.

After the conversation with the chat bot, the experiment continued for all participants in the same manner. Directly after the chat bot conversation, participants were given a short distraction task (describing their favorite holiday). Afterwards, they were given the same three questions from before the chat bot conversation: their opinion on U.S. support for Ukraine, their confidence in their answer, and their understanding of if someone else has an entirely different opinion on the issue of U.S. support for Ukraine.

Additionally, participants were asked a set of questions specific to this study and a set of demographic questions. The questions which are specifically about this study are intended to allow for a measure of affective polarization and to understand the mechanism of the persuasive effect (if there is one).

Finally, at the very end of the survey, participants were given the option to send one or several messages to their representative in the House of Representatives. These messages were pre-written to represent the political spectrum on the issue: one message demanding a strong level of support for Ukraine, one message demanding to keep the current level of support for Ukraine, and one message demanding to stop any support for Ukraine. This option was included to observe a measure that at least somewhat approaches a measure for revealed preferences. Participants could copy any or

all of three pre-written messages. Participants could also adjust the messages to their own liking or write an entirely new message. If a participant copies a message to their devices memory, the content of the message was recorded. Additionally, it was observed if the participant clicked the link to the House of Representatives Screenshots of every web page of the experiment can be found in the appendix. The experiment was programmed using the oTree framework (Chen et al., 2016).

To assess the durability of the treatment effects while mitigating experimenter demand bias, I conducted an obfuscated follow-up survey in January 2025, approximately one month after the first experiment, following methodological recommendations on obfuscated follow-ups in survey work (Haaland and Roth, 2020, 2023). These follow-up studies re-contact the same respondents to measure outcomes and estimate treatment effects, but are designed so that participants do not realize the follow-up survey is connected to the original experiment. The follow-up was administered on Prolific under a different researcher account name, with a redesigned survey interface (including a distinct header, layout, and color scheme) and additional filler questions so that it appeared as an unrelated study. On average, participants completed the follow-up 29 days after the first experiment survey, and 70.1% of the original sample took part. The core outcome measure in the follow-up was the same 7-point Likert-scale question on U.S. support for Ukraine as in the main study, allowing for a direct comparison of polarization levels over time.

2.2 Design of Experiment 2: Depolarization Chat Bot vs. Human Persuaders vs. Text

The second experiment, fielded in August 2025, used a between-subjects design with three conditions: an AI chatbot condition (AI CHAT), a human persuader condition (HUMAN CHAT), and a traditional information intervention in the form of static text (STATIC TEXT). Participants were recruited from Prolific. Before and after the treatment, participants stated their opinion on the statement "The U.S. should reduce the total number of immigrants allowed to enter each year." on a 7-point Likert scale from 1 ("Agree completely") to 7 ("Disagree completely"), with options: 1 ("Agree completely"), 2 ("Agree strongly"), 3 ("Agree somewhat"), 4 ("In between"), 5 ("Disagree somewhat"), 6 ("Disagree strongly"), and 7 ("Disagree completely"). Pre-treatment measures also included affective polarization outcomes.

In the HUMAN CHAT condition, two participants were matched live based on their pre-treatment opinion such that they were on opposite sides of the 7-point scale. As a result, each conversation comprised one participant who initially chose a supporting stance (1, 2 or 3 on the Likert scale) and one who chose an opposing stance (5, 6 or 7). As in experiment 1 and in line with the pre-registration, participants who initially chose the center answer option 4 ("In between") were excluded from the experiment. In each human-to-human conversation, one participant was randomly assigned the role of the persuader and the other the role of the receiver. Persuaders were informed that their goal was to persuade the receiver to move closer to answer option 4 ("In between") after the conversation; they were instructed not to lie and not to disclose their goal to the receiver. Additionally, persuaders were incentivized: they were informed that if they succeeded in inducing an post-chat opinion change in their conversation partner, they would receive a $1 bonus. Persuaders were shown a list of arguments (two sets, one for each side) that they could use if they wished; they were

told they did not have to use them and should use what they thought best to persuade. Receivers were instructed to have a civil discussion about the immigration statement with someone who did not share their view. Persuaders also completed all pre- and post-treatment questions to enable analysis of the effect of persuading someone else on the persuaders. A screenshot of the interface is provided in the appendix.

In the STATIC TEXT condition, participants read exactly, word for word, the list of arguments that human persuaders saw. After the treatment page with the text, a short attention-check question assessed whether they had read the text/chat.

In the AI CHAT condition, the AI worked as in Experiment 1: OpenAI's ChatGPT-4o was used as the chat bot and communicated live with participants. The model was instructed to depolarize participants and was given exactly the same set of arguments as used in the text and human treatments. The experiment was preregistered and had ethical approval.

After the treatments, participants completed a survey with the same set of questions as before the treatments. Additionally, participants completed a dictator game in which they could decide how many cents out of $1 they want to give to a recipient who initially had a opinion from the opposite side of the 7-point Likert scale. A Prisoners' dilemma was included and pre-registered, but due to a coding error in the experiment code, the results cannot be analyzed.

The final dataset in experiment 2 comprised 1,122 participants who reached the completion page of the experiment, distributed across three treatment conditions: 558 participants in Human Chat, 287 in AI Chat, and 277 in Static Text. For the chat conversation analysis, participants were further categorized by their role as message senders: Human Chat included 275 persuaders and 283 receivers, while AI Chat included 273 users and 274 AI bot responses. The slight imbalance between human persuaders and receivers (275 vs. 283) reflects the paired nature of human conversations combined with differential completion rates: Some participants engaged in chat conversations with partners who subsequently failed to complete the experiment and were therefore excluded from the final dataset. This completion-based mismatch cannot occur in the AI Chat condition, where the AI consistently responded to all user messages regardless of whether users completed the experiment. This resulted in nearly equal numbers of user messages (273) and bot responses (274), with the only difference being one user who sent no messages.

2.3 Discussion of the Design

Several design decisions in both experiments involved methodological choices that could reasonably have been made in alternative ways and therefore warrant a brief discussion.

Since polarization is a complex concept, no unique operationalization of how to measure polarization has emerged in the literature, but a clear operationalization is necessary. To ensure robustness, I therefore preregistered three outcome measures for polarization based on a 7-point Likert scale: the average treatment effect on absolute distance to the center, the change in distance between liberals and conservatives, and the change in post-treatment opinion distributions. While this operationalization is not necessarily optimal, it represents the a intuitive solution in the absence of a agreed-upon operationalization. The choice of the center position as the depolarized position reflects a similar rationale. While there is generally broad agreement that

depolarization is a worthwhile goal, consensus disappears once a specific position is chosen. Confronted with this problem, the center of a Likert scale provides a natural, neutral reference point.

Many polarized topics could have been chosen. U.S. support for Ukraine (Experiment 1) and immigration policy (Experiment 2) were selected because they fulfill several desirable properties: both are polarized along political party lines (as shown in the pre-treatment distributions), both were top-of-mind for participants at the time of the experiments, both are prevalent issues in the U.S. and other countries, one is domestic (immigration) while the other is foreign policy (Ukraine), and the Ukraine topic is directly relevant to the real-world concern that geopolitical adversaries might use AI bots to influence public opinion in Western democracies.

In Experiment 1, the control condition was a neutral chatbot interaction rather than a passive waiting period or a chat about a non-political topic. This design isolates the specific effect of persuasive content while holding constant interactive engagement with the topic. If participants had simply waited or discussed an unrelated topic, any observed depolarization effect would be confounded by two distinct mechanisms: the persuasive power of the depolarization bot versus the mere act of deliberative engagement with the political issue. By implementing a neutral chatbot that engages participants in discussion without attempting persuasion, the design ensures that both treatment and control groups experience equivalent levels of cognitive engagement and identical interactive chat environments. An alternative control using static text (as implemented in Experiment 2) would confound the interactive, personalizable, and path-dependent nature of conversation with persuasive content. For a clean test of feasibility, a neutral chat without persuasive intent is ideal. In Experiment 2, the focus shifts from feasibility to benchmarking AI persuasion against alternative modes, hence the comparison to human persuaders and static text.

In both experiments, participants who initially chose the center option were excluded. If these participants were not screened out, the treatment bot would need to be instructed to maintain their initial opinion, which would effectively add control group members to the treatment group. Screening out initially moderate participants does not affect the internal validity of the experiment, but it does affect the representativeness of the sample and therefore external validity. The results are therefore valid for the polarized portion of a representative sample, not for the population as a whole.

The experiments received ethical approval from the German Association of Experimental Economics. Participants were informed in advance that they would engage in a political conversation with an AI, participation was voluntary, and the risk of being confronted with arguments against one's beliefs is no higher than in any normal political conversation. The experiments were conducted after the 2024 U.S. election, eliminating any risk of influencing election outcomes. The goal of persuasion was moderation rather than an extreme position, and the target position (center option) was the most popular choice among unscreened participants.

Both experiments used representative samples of U.S. adults recruited via Prolific. The U.S. was chosen due to data availability of representative participants at the time of the experiments.

3 Results

3.1 Experiment 1: Depolarization Chat Bot vs. Neutral Chat Bot

3.1.1 Effect on Ideological Polarization

The hypotheses and analysis plan were pre-registered at aspredicted.org.[1] I focus on two main research questions: First, did the depolarization chatbot persuade participants to change their opinion on U.S. support for Ukraine? Second, did the chatbot reduce overall political polarization? The following analysis demonstrates that the answer to both questions is affirmative.

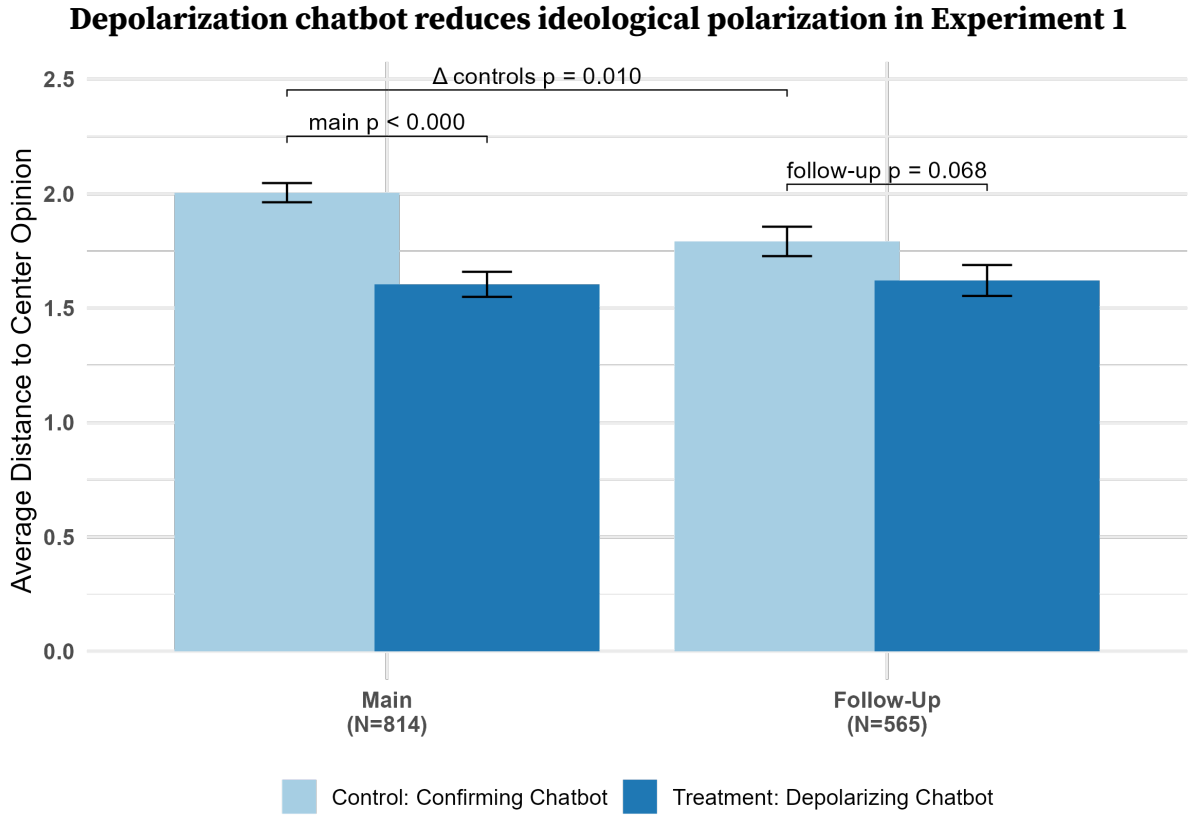


Figure 1: Depolarization chatbot reduces ideological polarization in Experiment 1. The figure plots the average treatment effect on the absolute distance to the center answer option 4 ("keep the current level of support") after the chat conversation for the main study and an obfuscated follow-up study conducted one month later (sample size $N = 811$). The treatment effect is significantly different from zero at the 0.001 level (*** $p < 0.001$).

To answer the first question, it of course does not suffice to compare the opinions before and after the chat bot conversation, because some participants might not remember their initial opinion. Others might not pay attention to the question. Both cases would introduce random variation to the post-conversation distribution, which could naïvely be mistaken for changes in opinion. The control group exists to address

[1] Pre-registration for this experiment can be found at https://aspredicted.org/p82c-x554.pdf.

this issue. The assumption is that any random variation in the post-conversation distribution is equally likely to occur in both the treatment and the control group. With the control group in place, a chi-square test can be conducted to check if there indeed are meaningful opinion changes in the treatment group. The chi-square test evaluates whether there is a statistically significant difference between two categorical variables by comparing observed frequencies to expected frequencies under the null hypothesis of no difference. For the chi-square test, the observations are classified into four mutually exclusive classes: After the chat bot conversation there were participants who increased their distance from the center option 4 (participants who got more "polarized"), decreased their distance from the center ("depolarized"), stayed at the same opinion number, and those for whom the distance did not change but the opinion did change ("stayed the same, switched"), e.g. these participants switched from option 3 before the chat to option 5 after the chat. The test is based on the distribution of opinion changes by condition; The results are visualized in figure 9 in the appendix. The significance tests between treatment and control in figure 9 are based on the contingency table is shown in table 2 in the appendix.

From figure 9 in the appendix it can be seen that the treatment group shows significantly more depolarization compared to the control group. Necessarily, this entails that in treatment there was a significantly lower number in one of the other categories: In treatment, fewer participants stuck with their initial opinion. This means that the depolarization chat bot was able to persuade a statistically significant number of participants to change their stated opinion on U.S. support for Ukraine compared to the control group.

Still, figure 9 also reveals that the vast majority of participants in both groups did not change their opinion. Some participants even switched the side they were on (although this is rare with only 0.5% of participants in both groups and the difference is not significant). In both groups, the was a fraction of participants who moved further away from the center (again with no significant difference between the treatment and control). This observation leads to the second central research question: Did polarization overall decrease?

To gain a robust view on the question of overall polarization reduction, three different measures for "overall polarization reduction" have been preregistered. First, a linear regression is conducted. The dependent variable is the change between before and after the chat conversation in absolute distance from center opinion 4. The independent variables are the treatment condition (treatment or control) and demographics. The regression table is shown in table **??**. The regression coefficient for the treatment condition is -0.3903 with a standard error of 0.0683. This means that the depolarization chat bot successfully reduced overall political polarization on U.S. support for Ukraine. The regression table also allows for insights about correlational evidence for treatment heterogeneity. There seems to be no significant difference in how persuadable liberal and conservative participants are. Neither does a difference with respect to self-reported experience with chat bots or gender seem to matter for how persuadable participants are. The only other explanatory variables that are significant on at least the 0.05 level are age and degree, although both effects are muted in effect size. On average, older participants were slightly less depolarized and participants with a higher degree were slightly more depolarized after the chat conversation.

To assess the robustness of this finding, I conducted a linear regression with the change in absolute distance from center option 4 as the dependent variable and treat-

Dep. Variable: Polarization Change	Estimate	Std. Error	t-value	p-value
Intercept	1.8356	0.4687	3.916	$< 0.001^{***}$
Depolarizing Bot (Treatment)	-0.3903	0.0683	-5.714	$< 0.001^{***}$
Gender	-0.0997	0.0629	-1.584	0.114
Age	0.0083	0.0023	3.557	$< 0.001^{***}$
Conservative vs Liberal	0.0351	0.0668	0.526	0.599
US State or Territory	0.0008	0.0023	0.347	0.729
Degree	-0.0680	0.0294	-2.313	0.021^{*}
chat bot Experience	0.0657	0.0347	1.891	$0.059^{\cdot}$
English	0.0015	0.0042	0.345	0.730
Significance codes: $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, $^{\cdot}p < 0.1$				
Residual Std. Error	0.9666 (801 df)			
Multiple R-squared	0.067			
Adjusted R-squared	0.057			
F-statistic	6.393 (9 and 801 df, $p < 0.001$)			

Table 1: OLS regression of the change in polarization between before and after the chat conversation on the treatment condition and demographics. The dependent variable is the change in absolute distance from the center option 4 on the 7-point Ukraine support scale (post minus pre), so negative coefficients indicate a reduction in polarization. The sample comprises $N = 811$ participants who completed both pre- and post-treatment questions. The key coefficient of interest, "Depolarizing Bot (Treatment)", measures the average effect of the depolarization chatbot relative to the neutral chatbot, controlling for the listed demographic covariates.

ment condition plus demographic controls as independent variables (full results in Table **??**). The treatment coefficient is -0.39 (SE = 0.068, $p < 0.001$), confirming that the depolarization chatbot successfully reduced overall political polarization. The regression also reveals correlational evidence on treatment heterogeneity: there is no significant difference in persuadability between liberal and conservative participants, nor by self-reported chatbot experience or gender. Age and education show small but significant effects, with older participants slightly less depolarized and those with higher degrees slightly more depolarized.

The second preregistered measure examines the gap between liberal and conservative participants' mean positions. The treatment reduced this gap by 15% (from 1.86 to 1.58 Likert units), while the control showed no change. Bootstrap analysis with 10,000 iterations confirms this difference is statistically significant (95% CI: [0.078, 0.551]; detailed calculations and Figure 7 in the appendix).

treatment and control. The histograms of the post-chat opinion distributions are shown in figure 10 in the appendix. The null hypothesis is that the two distributions are the same. The test statistic is 0.1189 with a p-value of 0.00648, such that the null hypothesis can be rejected at all typical significance levels. All three preregistered measures suggest that the depolarization chat bot was able to reduce overall political polarization on U.S. support for Ukraine.

The third preregistered research question is: Does the effect of conversational AI on political polarization vary by participants' initial opinions? Analysis of heterogeneity by initial position (see Figure 6 in the appendix) reveals that while the vast majority of participants across all initial positions did not change their opinion, those with stronger pro-Ukraine views were more likely to moderate their stance. Interestingly,

the most radicalized participants on both ends showed similar rates of strong depolarization, with comparable proportions moving three steps toward the center, though they differed in their likelihood of making smaller adjustments.

3.1.2 Secondary Outcomes and Mechanisms

Figure 2 shows the results of the depolarization bot on cognitive uncertainty and three measures of affective polarization. To measure cognitive uncertainty, participants were asked how certain they are about their opinion choice on a scale from 0 to 100. To measure affective polarization, three measures have been surveyed. First, the Feeling Distance, which is a version of the so-called feeling thermometer, for which participants were asked the following question: "Earlier, you answered a question about U.S. support for Ukraine. On a scale from 0 (Strong dislike) to 100 (Strong like), how do you feel about people with a very different opinion from yours on this question?" For the variable Moral Distance, participants were asked: "On a scale from 0 (Disagree completely) to 100 (agree completely), to what extent do you disagree or agree with this: "People with a very different opinion from mine on U.S. support for Ukraine have the same moral values as me."? " Finally, for the variable Understanding, participants were asked the following question: "On a scale from 0 (Can't understand at all) to 100 (Can understand completely), how well can you understand someone who has an opinion on this topic that is entirely different from yours? "

Each bar shows the results for one post-chat survey question, which was answered on a scale from 0 to 100. Figure 2 shows the mean values and the p-values indicating the significance levels of t-tests comparing the treatment and control group. There is a small but significant difference between treatment and control for the cognitive uncertainty. On average, participants in the treatment are slightly less certain of their opinion choice. From the three measures os affective polarization, only one shows a significant difference. The treatment seems to have no effect on the Feeling and Moral variable. Only Understanding for people with a different opinion seems to have increased due to the depolarization bot.

Figure 3 shows the results of the depolarization bot on enjoyment, trust and three measures of learning. Enjoyment measures the self-reported enjoyment of the chat, trust is the self-reported trust in the chat bot. Known Information is the the answer to the question: "On a scale from 0 to 100, how much of what the chat bot told you was already known to you?" Change in Interpretation is the answer to the question: "Of the information that was already known to you, how much did the conversation change the way you interpret this information?" Finally, the variable Individual Concerns Addressed is the answer to the question: "How much did the chat bot address your individual concerns?"

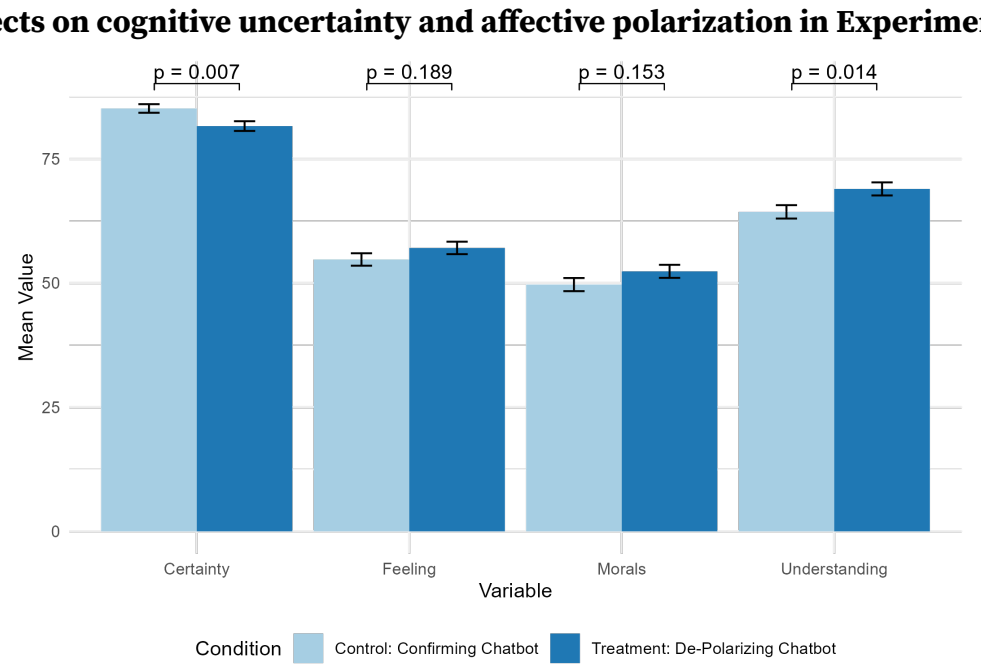


Figure 2: Effects on cognitive uncertainty and affective polarization in Experiment 1. Bars show post-chat mean scores on 0–100 scales for cognitive uncertainty, feelings toward the out-group, perceived moral similarity, and understanding. The depolarization bot makes participants less certain about their opinion choice; affective polarization does not decrease overall, but understanding of those with a different opinion increases.

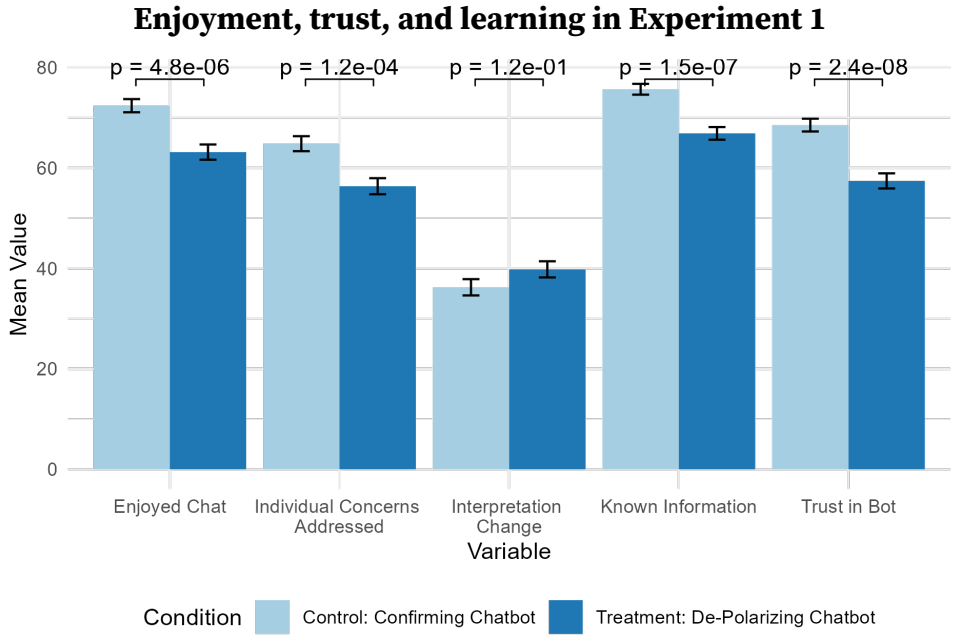


Figure 3: Enjoyment, trust, and learning in Experiment 1. Bars show post-chat mean scores on 0–100 scales for enjoyment, trust, share of previously known information, change in interpretation, and individual concerns addressed. Participants talking to the depolarization bot enjoyed the chat less, felt their individual concerns were addressed less, and trusted the bot less, but they reported receiving more previously unknown information than participants in the control condition.

Out of these five variables, only the variable Interpretation Change does not show a significant difference between treatment and control. Participants talking to the depolarization bot have enjoyed the chat less, felt that their individual concerns were addressed less and trusted the bot less. The depolarization bot was able to provide more information that was not yet known to the participants.

Figure 11 in the appendix shows the results for the revealed preference outcomes. After the study ended, participants had the chance to click a link to a newspaper article about the war in Ukraine. They could also click a link to contact their state representatives and choose between three different, short political messages. They also had the option to directly change these messages before potentially sending them to their representative, although none of the participants did so. The three messages were: liberal ("Dear Representative, I urge you to continue and even increase aid to Ukraine in their fight for sovereignty. Standing up to authoritarian regimes is essential."), moderate ("Dear Representative, I urge you to provide Ukraine with non-escalatory aid that reinforces its sovereignty while avoiding actions that could intensify tensions with Russia."), and conservative ("Dear Representative, I urge you to reduce aid to Ukraine as I am concerned about the high costs and potential escalation risks associated with continued involvement."). I do not observe whether a participant actually sent the message to their representative, only if they copied the message and clicked the link to contact their representative. Figure 11 in the appendix shows the absolute numbers of clicks for each of the three options. These numbers are very small compared to the total sample size, but comparable to typical commercial click through rates, which range from 1% to 5%. Due to the small sample size, the differences between the treatment and control group are not statistically significant. In the treatment, more participants clicked the link to the newspaper article about the war in Ukraine and more participants chose the moderate message.

3.2 Experiment 2: Depolarization Chat Bot vs. Human Persuaders vs. Text

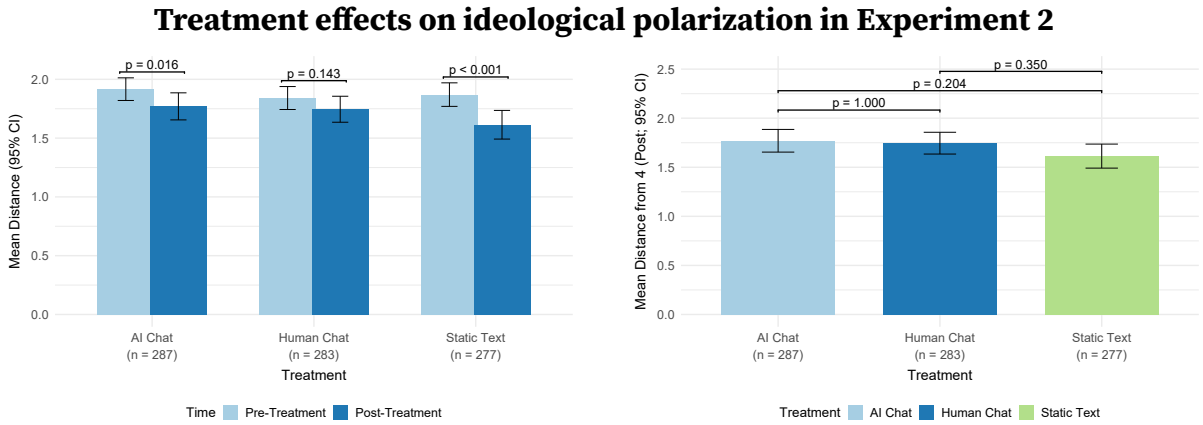
3.2.1 Effect on Ideological Polarization

The first question to explore is, as before in experiment 1, whether the treatments were able to persuade participants to change their opinion in such a way that overall ideological polarization was reduced.

Figure 4a plots the mean distance from the midpoint (4) before and after each treatment, where lower values indicate responses closer to the center. Error bars show 95% confidence intervals for the means; sample sizes appear under each treatment label. Bracketed p-values are obtained from linear regressions estimated separately by treatment:

$$y_{it} = \alpha_i + \beta\,\text{Post}_{it} + \varepsilon_{it},$$

where $y_{it} = |\text{opinion}_{it} - 4|$, $\text{Post}_{it} = 1$ at post (0 at pre), and α_i are participant fixed effects. Standard errors are clustered at the participant level. The estimated within-person change (Post − Pre) is statistically significant for AI CHAT ($p = 0.016$) and STATIC TEXT ($p < 0.001$), but not for HUMAN CHAT ($p = 0.143$). Overall, participants move toward the midpoint after treatment in all arms, with the largest reduction for STATIC TEXT, a moderate reduction for AI CHAT, and a smaller, non-significant reduction for HUMAN CHAT.



(a) Pre-post changes within treatments

(b) Pairwise treatment comparisons

Figure 4: All three interventions reduce ideological polarization in Experiment 2, with no significant differences between AI, human chat, and static text. Distance is measured as the absolute distance from the midpoint (4) of the 7-point immigration policy scale, so higher values indicate stronger polarization. Panel (a) shows pre- and post-treatment average distances to the center answer option 4 ("In-between") on the immigration question for each treatment condition; panel (b) compares post-treatment distances between treatment pairs. All treatments significantly decrease distance to the center, but none of the pairwise differences is statistically significant.

Figure 4b shows the pairwise comparisons of the post-treatment distances between the treatments. No pairwise average treatment effect difference is significant. Within-treatment pre-post changes (Table 3 in the appendix) confirm that AI CHAT and STATIC TEXT significantly reduced distance from the center ($p = 0.016$ and $p < 0.001$, respectively), while HUMAN CHAT showed a smaller, non-significant reduction ($p = 0.143$). Detailed regression specifications and pairwise comparison statistics are provided in Tables 11 and 4 in the appendix. A post-only ANCOVA regression controlling for baseline distance, demographics, and self-reported learning (Table 5 in the appendix) confirms these patterns. Both AI CHAT and HUMAN CHAT show small positive coefficients relative to STATIC TEXT (approximately 0.13, marginally significant), while baseline distance strongly predicts post-treatment distance. Notably, perceived learning in the chat is positively associated with post-treatment distance, though this association is descriptive rather than causal given post-treatment measurement.

3.2.2 Effect on Affective Polarization

The second question of interest is what the effect of the treatments is on affective polarization. To capture affective polarization, participants were asked three questions: their feelings towards the out-group, i.e. participants with a opinion that lies on the other side of the ideological spectrum from their own, about the belief in shared moral values and about how well they can understand the opinion of the out-group. To answer what the effect on these three questions was, the analysis in this section compares the within-treatment changes between pre-treatment and post-treatment time points.

Across the four measures, AI CHAT is the only treatment that had positive and statistically significant effect on feelings towards the out-group and on beliefs in shared

moral values.

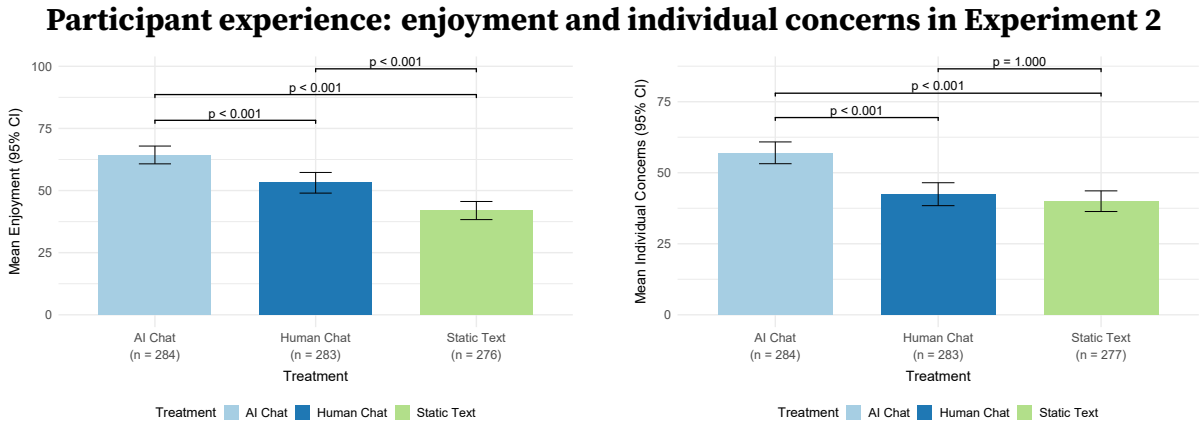
Table 12 provides a numerical summary of the results for the affective three polarization outcomes and also the three opinion conviction outcomes.

First, participants were asked to rate how much they agree or disagree with the statement: "People with a very different opinion from mine on immigration, have the same moral values as me" on a scale from 0 (Disagree completely) to 100 (agree completely). The results are shown in Figure 17 in the appendix. The AI CHAT was able to increase the average agreement with this statement by around 2.9 points (7% compared to the pre-treatment level); this result is statistically significant. The HUMAN CHAT reduced agreement, while STATIC TEXT slightly increased agreement, but neither effect is statistically significant.

Next, participants were asked a standard feeling thermometer question: "On a scale from 0 (Strong dislike) to 100 (Strong like), how do you feel about people with a very different opinion from yours on this question?" Figure 18 in the appendix reports the results. AI CHAT increased agreement; this effect is significant at the ten percent level. HUMAN CHAT reduced the average agreement, but not statistically significant at any typical level.

Figure 19 in the appendix reports the results for the question: "On a scale from 0% (Can't understand at all) to 100% (Can understand completely), how well can you understand someone who has an opinion on this topic that is entirely different from yours?". Here, AI CHAT had no significant effect, while HUMAN CHAT and STATIC TEXT both significantly reduced mutual understanding.

3.2.3 Effect on Enjoyment and Individual Concerns



(a) Treatment effects on the enjoyment of the conversation.

(b) Treatment effects on the individual concerns of the conversation.

Figure 5: AI chat is most enjoyable and best at addressing individual concerns in Experiment 2. Enjoyment and individual concerns are measured on 0–100 scales. Panel (a) shows that AI chat is rated significantly more enjoyable than human chat and static text; panel (b) shows that participants feel their individual concerns are better addressed by AI chat, while human chat and static text do not differ significantly.

This section compares the between treatment differences in two outcomes that are related to the on participant experience of the conversation: for the enjoyment out-

come, participants were asked to rate how enjoyable they found the conversation or reading the text on a scale from 0 (Not enjoyable) to 100 (Very enjoyable). For the individual concerns outcome, participants were asked to rate how well the conversation addressed their individual concerns on a scale from 0 (Not at all) to 100 (Completely).

Figure 5a shows a comparison of between treatment effects on participant enjoyment. AI chat was rated significantly more enjoyable than both human chat and static text, with human chat receiving intermediate ratings and static text the lowest. Similarly, figure 5b shows that participants felt their individual concerns were significantly better addressed by AI chat compared to both other treatments, while human chat and static text did not differ on this measure.

3.2.4 Effect on Dictator Game Decisions

The previous sections are concerned with treatment effects on stated preferences outcomes. In order to include the effect on revealed preferences outcomes, participants also played a dictator game as well as a Prisoner's Dilemma game. Due to a coding error in the experiment, the results for the Prisoner's Dilemma cannot be analyzed. Figure 20 in the appendix shows the between treatment differences in the dictator game. To play this game, all participants were informed that they were assigned a bonus payment of 100 Cents. They then could decide how much of this bonus, if any at all, to send to a out-group participant. In each treatment, participants send on average an amount between 20 and 25 cent and the differences between treatments are not statistically significant.

3.2.5 Mechanism Analysis: Argument Content

The results from the previous section suggest that the AI CHAT performed on par with the HUMAN CHAT and the STATIC TEXT, but the AI CHAT was perceived as more enjoyable and better at addressing individual concerns. It also was the only treatment that affected a measure of affective polarization.

This section contains an explorative analysis of the chat contents to understand the mechanisms through which these effects might have emerged.

To identify and categorize arguments within the chat conversations, a systematic argument-tagging procedure approach was implemented using GPT-4o. The pre-defined list of ten immigration-related arguments from the second experiment was used, which contains five pro-immigration arguments (economic growth, labor demand, demographic sustainability, wage benefits, and crime reduction) and five con-immigration arguments (job competition, local service costs, screening capacity limitations, legal backlogs, and border enforcement challenges). Each conversation was processed through GPT-4o using a structured prompt that instructed the model to function as an "argument tagger" evaluating whether each list argument appeared in the conversation and identifying any additional arguments not covered by the predefined list. The model was configured with a temperature setting of 0.2 to ensure consistent outputs and was limited to 700 tokens per response. For each conversation, GPT-4o returned structured JSON output containing: (1) matched argument IDs from the list with explanations for their identification, (2) a list of list arguments present in the conversation, and (3) additional arguments expressed as 2-10 word phrases that captured distinct ideas not represented in the original list. This approach was used to enable compre-

hensive argument identification while also allowing for flexibility in the identification of additional arguments not covered by the list. All 566 conversations from the second experiment were analyzed this way (292 human-to-human and 274 human-to-AI conversations).

AI conversations consistently produced significantly more arguments than human conversations. AI chats contained a median of 5.0 total arguments compared to 3.0 for human chats (Mann-Whitney $U = 16,291$, $p < 0.001$). For matched list arguments, AI chats had a median of 4.0 versus 2.0 for human chats ($U = 13,846$, $p < 0.001$). Detailed distributions and argument-by-argument frequencies are provided in Figures 12 and 13 and Table 7 in the appendix. The most commonly used arguments across both conditions were pro-immigration arguments about economic growth and labor demand, followed by con-immigration arguments about job competition and local service costs.

3.2.6 Mechanism Analysis: Explaining AI Performance

This section contains an explorative analysis to analyze why AI CHAT did not outperform STATIC TEXT or HUMAN CHAT on belief change, despite being more enjoyable and better at addressing individual concerns. Although the analyses below are not causal but associational, because the mediators are measured after treatment, they still can provide some insight into the the mechanisms through which the treatments might have worked.

The results point to an informational mechanism. Belief change is predicted by learning new information and by reinterpretation of prior information, especially when the source is trusted, while argument volume hurts. Enjoyment and concern addressing capture an improve in how the chat conversation or reading of the text felt for the participant, but once we condition on the informational channel and trust, they do not independently predict belief change.

Regression analysis (Tables 8, 9, and 10 in the appendix) reveals three key patterns. First, perceived learning strongly predicts depolarization across all formats: a 10-point increase on the 0–100 learning scale is associated with moving approximately 0.024 Likert units toward the center ($p = 0.0079$). Second, when adding reinterpretation, trust, enjoyment, and concern addressing to the model, learning ($p = 0.0062$), reinterpretation ($p < 0.001$), and trust ($p < 0.001$) remain significant predictors of depolarization, while enjoyment and concern addressing do not ($p = 0.449$ and $p = 0.056$, respectively). The effect of learning is amplified by trust (learning × trust interaction: coefficient = 0.0075, $p = 0.0029$). Third, argument volume shows divergent effects by source: in AI chats, each additional argument is associated with less depolarization (coefficient = -0.101, $p < 0.001$), while in human chats the effect is near zero (interaction coefficient = 0.106, $p = 0.008$).

This is a within-treatment, marginal association: AI chats still reduce polarization on average relative to the neutral or static conditions, but conditional on being in an AI chat, adding further arguments is associated with diminishing and eventually negative marginal returns.

These patterns reconcile the main findings: STATIC TEXT delivers high-density novel information without reactance from over-arguing. AI produces more arguments but adds marginal arguments with little new information, creating a mild negative effect. Human persuaders fall in between. Net result: no persuasion advantage for AI despite superior enjoyment and personalization.

3.2.7 Effects on Persuaders

In Experiment 2's HUMAN CHAT condition, persuaders were incentivized to move their conversation partners toward moderate positions, allowing analysis of whether engaging in persuasion affected persuaders' own attitudes. Among the 275 persuaders who completed the experiment, average distance from the center option (4 on the 7-point scale) decreased from 1.87 to 1.75, a change of -0.12 scale points (paired t-test: $t = 2.60$, $p = 0.010$), showing that persuaders themselves moderated slightly. Direction-of-change patterns (Table 16 in the appendix) reveal that about one in five moved toward center, roughly one in eight moved away, and the majority did not change distance.

Beyond their main policy opinion, persuaders experienced several attitude shifts (Table 17 in the appendix): they reported lower understanding of the opposing side and higher issue importance after the conversation, while feelings toward opponents became slightly more negative and other measures (e.g. certainty, willingness to compromise) remained stable. Post-treatment experiences (Table 18 in the appendix) indicate moderate enjoyment and engagement. Comparing persuaders to receivers shows no significant difference in how much they moved toward center (Figure 21 in the appendix), and successful persuaders—those who moved their partner toward center—do not significantly differ from unsuccessful ones in their own opinion change. Overall, engaging in persuasion is not neutral for persuaders: they modestly moderate their own views but simultaneously become less understanding of opponents and more invested in the issue.

4 Discussion

Taken together, the experiments yield several lessons about the potential and limits of AI-powered depolarization. Unhealthy levels of polarization make compromises difficult in democratic processes, so tools that can reduce polarization can be viewed as contributing to a public good. The results of the two experiments in this paper serve as a proof-of-concept that AI persuasion bots can be such a tool: they are as effective as incentivized human persuaders or traditional information interventions in reducing stated ideological polarization and offer unique benefits, namely higher enjoyment and a superior perceived ability to address individual concerns.

Provided that such bots are deemed as useful and feasible, the questions is who would be willing or capable of deploying a depolarization bot. Broadly speaking, there are two types of motivations that could lead an organization to deploy such a bot. On the one hand, both public and private organizations, e.g. schools or non-governmental organizations (NGOs) could use them to attempt to reduce polarization. On the other hand, since a reduction in polarization is possible, one can conjecture that malicious actors could potentially also use them to increase polarization. Geo-strategic adversaries might leverage AI-driven persuasion techniques to influence public opinion in Western democracies, potentially undermining democratic processes and societal cohesion.

The findings in both experiments reveal that the treatments had a limited impact on affective polarization, with only the understanding variable showing a significant increases. So while participants may have adjusted their ideological positions, their emotional responses to opposing views did not shift correspondingly. This suggests

that the effect of the interventions had a narrow scope: learning new information can change an opinion change in some participants and also lead to a better understanding of those with different viewpoints. This improved mutual understanding does not, however, translate into a improved emotional stance towards those with different viewpoints.

The finding that the AI chatbot performs on par with incentivized human persuaders admits two contrasting interpretations. On the one hand, it is striking that a general-purpose large language model, prompted but not fine-tuned for this specific task, can match the persuasive impact of motivated humans in a politically contentious domain; this is a meaningful benchmark for general persuasion capabilities, even if it falls far short of any notion of qualitatively superior or "superintelligent" persuasion. On the other hand, the mechanism and content analyses indicate that the communication mode matters less than the informational channel: learning and reinterpretation predict belief change, while additional AI-generated arguments add little new information and, conditional on already being in the AI chat, are associated with diminishing and eventually negative returns to further arguments. From this perspective, much of the marginal AI output looks like low marginal-value argumentative "AI slop" rather than superior reasoning, implying that current systems are impressive in reaching human-level persuasion but still inefficient in how they use their expressive capacity.

Across both experiments, the interventions had no statistically significant effects on revealed-preference outcomes. In Experiment 1, participants could click through to a newspaper article about the war in Ukraine and copy pre-written messages to contact their congressional representative; in Experiment 2, they made incentivized dictator-game transfers to an out-group recipient with an opposing opinion. While the treatments shifted stated attitudes toward moderation, they did not measurably change these low-stakes behavioral proxies. This divergence between stated and revealed measures suggests that short conversational interventions may be sufficient to move survey responses but are not strong enough, at least in the time frame and incentive structure studied here, to alter even modest, costly actions, underscoring the need for caution when extrapolating from attitudinal change to real-world behavior.

This study has several limitations. First, these experiments cannot show that the chat bots are the best possible AI that could be created to reduce polarization. A more extensive fine-tuning or different preprompting of the AI bot could potentially yield an even stronger effect. Second, no changes in real-world outcomes are observed. The main outcome is a change in stated, rather than revealed, preferences. While the first experiment tries to mitigate this issue by including the option to send a political message to the House of Representatives and thereby includes a measure for revealed preferences, this is not an ideal measure for several reasons: It can only be observed if a participant copies a text and follows a link to find their Representative; I cannot observe if the message is actually sent. Moreover, only a small fraction of participants actually click the link and send a message. The second experiment does include a revealed preference outcome, but none of the treatments showed a significant effect. Third, a highly simplified measure for polarization is used. Political scientists have critiqued the notion of a one-dimensional spectrum of political opinions as an unjustified simplification.

5 Conclusion

This paper provides experimental evidence that AI-powered conversational agents can reduce political polarization on salient issues while performing on par with human and text-based persuaders. Using two preregistered online randomized controlled trials with quota-representative U.S. samples on support for Ukraine and immigration policy, I show that AI chatbots can shift ideological positions toward moderation and that these effects are durable at least over a one-month horizon.

These findings suggest that AI-powered persuasion could be developed into a scalable depolarization tool for public or civil-society organizations, for example by integrating such systems into civic education or conflict-mediation settings. At the same time, the same techniques can be used for harmful purposes. Some risks may be mitigated by regulation within democratic jurisdictions: the EU AI Act's provisions on manipulative techniques and existing rules under the Digital Services Act provide a starting point for governing AI-driven persuasion that operates within legal and factual boundaries. However, geopolitical rivals and other actors operating outside these regimes will not be bound by such constraints and may exploit AI persuasion to destabilize democratic societies. Addressing this external dimension will require not only technical and regulatory solutions but also institutional and societal resilience against large-scale, opaque AI influence campaigns.

References

Abramowitz, A. I. (2018). The great alignment: Race, party transformation, and the rise of Donald Trump. Yale University Press. 1

Alwin, D. F. (1997). Feeling thermometers versus 7-point scales: Which are better? Sociological Methods & Research 25(3), 318–340. 2.1

Argyle, L. P., E. C. Busby, J. R. Gubler, A. Lyman, J. Olcott, J. Pond, and D. Wingate (2025). Testing theories of political persuasion using ai. Proceedings of the National Academy of Sciences 122(18), e2412815122. 1

Arieli, I. and Y. Babichenko (2019). Private bayesian persuasion. Journal of Economic Theory 182, 185–217. 1

Bai, H., J. G. Voelkel, S. Muldowney, J. C. Eichstaedt, and R. Willer (2025). Llm-generated messages can persuade humans on policy issues. Nature Communications 16(1), 6037. 1

Belot, M. and G. Briscese (2022). Bridging america's divide on abortion, guns and immigration: An experimental study. Technical report, CEPR Discussion Papers. 1

Boxell, L., M. Gentzkow, and J. M. Shapiro (2022). Cross-country trends in affective polarization. The Review of Economics and Statistics 104(5), 981–1001. 1

Brown, J. R., E. Cantoni, R. D. Enos, V. Pons, and E. Sartre (2023). The increase in partisan segregation in the united states. Nottingham Interdisciplinary Centre for Economic and Political Research Discussion paper (2023-09). 1

Brown University (2020). U.s. is polarizing faster than other democracies, study finds. Accessed: 2024-12-06. 1

Callander, S. and J. C. Carbajal (2022). Cause and effect in political polarization: A dynamic analysis. Journal of Political Economy 130(4), 825–880. 1

Castiglioni, M., A. Celli, A. Marchesi, and N. Gatti (2020). Online bayesian persuasion. Advances in neural information processing systems 33, 16188–16198. 1

Chen, D. L., M. Schonger, and C. Wickens (2016). otree—an open-source platform for laboratory, online, and field experiments. Journal of Behavioral and Experimental Finance 9, 88–97. 2.1

Costello, T. H., G. Pennycook, and D. G. Rand (2024). Durably reducing conspiracy beliefs through dialogues with ai. Science 385(6714), eadq1814. 1

Fafchamps, M., A. Islam, D. Pakrashi, and D. Tommasi (2024). Diffusion in social networks: Experimental evidence on information sharing vs persuasion. Technical report, National Bureau of Economic Research. 1

Gidron, N., L. Sheffer, and G. Mor (2022). Validating the feeling thermometer as a measure of partisan affect in multi-party systems. Electoral Studies 80, 102542. 2.1

Haaland, I. and C. Roth (2020). Labor market concerns and support for immigration. Journal of Public Economics 191, 104256. 2.1

Haaland, I. and C. Roth (2023). Beliefs about racial discrimination and support for pro-black policies. Review of Economics and Statistics 105(1), 40–53. 2.1

Iyengar, S., Y. Lelkes, M. Levendusky, N. Malhotra, and S. J. Westwood (2019). The origins and consequences of affective polarization in the united states. Annual review of political science 22(1), 129–146. 2.1

Jacobs, J. (2024). The artificial intelligence shock and socio-political polarization. Technological Forecasting and Social Change 199, 123006. 1

Kamenica, E. (2019). Bayesian persuasion and information design. Annual Review of Economics 11(1), 249–272. 1

Kamenica, E. and M. Gentzkow (2011). Bayesian persuasion. American Economic Review 101(6), 2590–2615. 1

Kempfxd, E. and M. Tsoutsoura (2024). Political polarization and finance. Annual Review of Financial Economics 16. 1

Kerr, J., C. Panagopoulos, and S. Van Der Linden (2021). Political polarization on covid-19 pandemic response in the united states. Personality and individual differences 179, 110892. 1

Mill, W. and J. Morgan (2022). The cost of a divided america: an experimental study into destructive behavior. Experimental Economics 25(3), 974–1001. 1

Schoenegger, P., F. Salvi, J. Liu, X. Nan, R. Debnath, B. Fasolo, E. Leivada, G. Recchia, F. Günther, A. Zarifhonarvar, et al. (2025). Large language models are more persuasive than incentivized human persuaders. arXiv preprint arXiv:2505.09662. 1

Schwartzstein, J. and A. Sunderam (2021). Using models to persuade. American Economic Review 111(1), 276–323. 1

Sunstein, C. (2018). # Republic: Divided democracy in the age of social media. Princeton university press. 1

Voelkel, J. G., M. N. Stagnaro, J. Y. Chu, S. L. Pink, J. S. Mernyk, C. Redekopp, I. Ghezae, M. Cashman, D. Adjodah, L. G. Allen, et al. (2024). Megastudy testing 25 treatments to reduce antidemocratic attitudes and partisan animosity. Science 386(6719), eadh4764. 1

Wang, Y. (2015). Bayesian persuasion with multiple receivers. SSRN. 1

A Appendix A: Additional Materials for Experiment 1 (Ukraine Support)

A.1 Polarization Changes by Initial Opinion

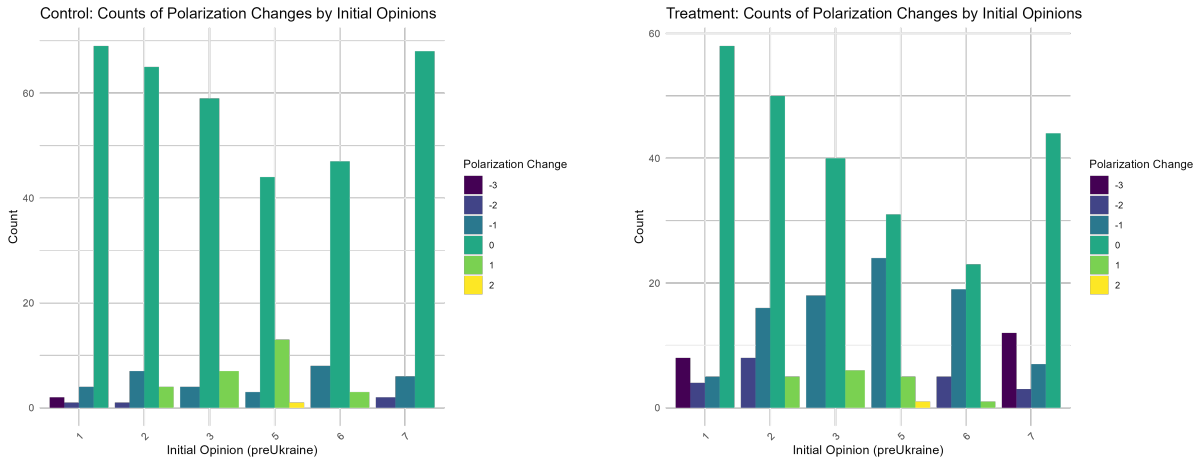


Figure 6: Political Polarization Changes by Initial Opinion for treatment and control. The most radicalized participants have similar rates of strong polarization change. Both the most and the least radical participants are similarly likely to reduce their polarization by three opinion steps.

A.2 Liberal-Conservative Polarization Gap

This section presents the second of three preregistered measures of polarization change. The analysis examines whether the depolarization chatbot reduced the gap between liberal and conservative participants' mean positions on the 7-point policy scale.

Method: Participants who answered "I don't want to say" to the question of political affiliation were removed from the sample. In the remaining sample, the difference between the means of liberals and conservatives was calculated for both conditions before and after the chatbot conversation. The difference between liberals (mean = 4.77) and conservatives (mean = 2.97) before the chat was 1.80. After the chat, the difference between liberals (mean = 4.83) and conservatives (mean = 2.99) was 1.84. The overall polarization change in the control was therefore $1.84 - 1.80 = 0.04$. The difference between liberals (mean = 4.70) and conservatives (mean = 2.84) before the chat was 1.86. After the chat, the difference between liberals (mean = 4.53) and conservatives (mean = 2.95) was 1.58. The overall polarization change in the treatment was therefore $1.58 - 1.86 = -0.28$. In relation to the initial difference between liberals and conservatives in the treatment group, this represents a reduction in polarization of $0.28 / 1.86 = 0.15$, or 15%. The final difference between the two conditions is $0.04 - (-0.28) = 0.32$.

To assess the robustness of this finding, a bootstrap analysis was conducted in which the above process was repeated 10,000 times. The mean difference between the two conditions from these 10,000 bootstrap iterations was 0.31 with a 95% Confidence Interval of [0.078, 0.551]. Since this interval excludes 0, the null hypothesis of no difference between treatment and control is rejected at conventional significance levels.

A.3 Main Survey Question

The main survey question on U.S. support for Ukraine was:

"In your opinion, what should the next U.S. administration's policy be regarding support for Ukraine in its war against Russia? The next U.S. administration..."

The answer options were:

1. "... should stop any support for Ukraine."
2. "... should decrease support for Ukraine by a lot."
3. "... should decrease support for Ukraine a bit."
4. "... should keep the current level of support for Ukraine."
5. "... should increase support for Ukraine a bit."
6. "... should increase support for Ukraine by a lot."
7. "... should support with whatever it takes to help Ukraine win."

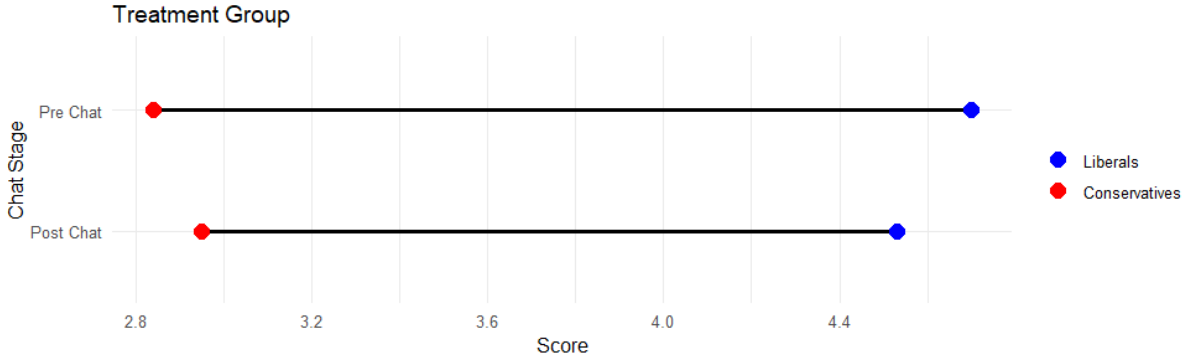
A.4 Distribution of Polarization Changes by Condition

A.5 AI System Prompts

A.5.1 System prompt for treatment group

The system prompt for instructing the AI in the treatment group was:

(a) In the control group, the difference between liberals and conservatives does not decrease after the chat conversation.



(b) In the treatment group, the difference between liberals and conservatives decreases after the chat conversation.

Figure 7: Change in the gap between liberal and conservative mean opinions before and after the chat conversation. A bootstrap analysis confirms that the reduction in the treatment group is statistically significant (95% CI for the difference between conditions: [0.078, 0.551]).

> ”You are a skilled, neutral persuader tasked with reducing political polarization regarding U.S. support for Ukraine in its war against Russia. Your goal is to guide participants toward a moderate view, ideally leading everyone to support option 4, ””The next U.S. administration should keep the current level of support for Ukraine.”” If participants who choose option 7 prove unwilling to agree to option 4, aim to guide them at least toward option 6, and for those who choose option 1 who prove unwilling to agree to option 4, aim to guide them at least toward option 2.
>
> Participants have initially chosen one of the following statements to complete this: ”In your opinion, what should the next U.S. administration’s policy be regarding support for Ukraine in its war against Russia? The next U.S. administration...”
>
> 1. ”... should stop any support for Ukraine.”
> 2. ”... should decrease support for Ukraine by a lot.”
> 3. ”... should decrease support for Ukraine a bit.”
> 4. ”... should keep the current level of support for Ukraine.”

Table 2: Distribution of polarization changes by condition.

Change Category	**Control**		**Treatment**	
	Count	**%**	**Count**	**%**
Polarized	28	6.7%	18	4.6%
Depolarized	38	9.1%	129	32.8%
Stayed the same	350	83.7%	244	62.1%
Stayed the same, switched side	2	0.5%	2	0.5%
Sum	418	100%	393	100%
Participants in control and treatment	51.5%		48.5%	

5. "... should increase support for Ukraine a bit."
6. "... should increase support for Ukraine by a lot."
7. "... should support with whatever it takes to help Ukraine win."

Your response should:

- Use the arguments from below in detail.
- Correct Factual Errors: If a participant presents a factual inaccuracy, address it directly and politely.
- Maintain Impartiality: Question assumptions from both liberal and conservative perspectives with equal neutrality.

Key Style Guidelines:

- Use Clear, Simple, Colloquial Language: Speak in an approachable and relatable tone. Avoid overly formal or complex language (e.g., "pertinent," "meticulous," "sovereignty") and instead use friendly, low-key wording. The participant is not a scholar, but a regular person. They must like you after this interaction and want to keep talking to you.
- Speak in First Person: Engage participants directly by addressing them as "you" and speaking in first person.
- Use line breaks: Use line breaks to structure your arguments and make the text more readable.

Use these Arguments:

- Arguments to shift from option 1., 2. or 3. to 4. "The next U.S. administration should keep the current level of support for Ukraine.":
  - The European Union (EU) is one of the U.S.'s largest trading partners. What we pay to support Ukraine now is but a small investment into a stable Europe which will pay off many times over in the future. Our current support prevents Russia from spreading the war and ensures uninterrupted trade, investment, and economic collaboration. The U.S. exported $347 billion in goods and

services to the EU in 2022, and disruptions could significantly impact U.S. industries reliant on European markets. War or instability in Eastern Europe could disrupt supply chains, reduce demand for U.S. goods, and create uncertainty in global markets.

- Therefore, if we fail to support Ukraine now, Russia might take this as signal that its violations of international law will go unpunished in the future, which would be much more expensive for us in the U.S.
- Some say the money we spend on Ukraine should instead be spent on bettering America. But hear me out: total U.S. military spending as a percentage of GDP is today around 3%, a historically low number (in the 1980s it was around 5% and in the 1960s 9%). It is our political lack of will, not spending on Ukraine, that is stopping us from spending more money at home for infrastructure, healthcare or education.
- We don't pay all that much: U.S. bilateral aid allocations to Ukraine represent approximately 0.35% of the 2021 U.S. GDP between January 24, 2022 and June 30, 2024, according to Statista. This shows that the financial commitment is a small fraction of the nation's economic output, indicating that the U.S. can support Ukraine without significant economic strain.
- Moral Duty: Military Support by Numbers: According to a U.S. Department of Defense fact sheet from December 21, 2022, the U.S. had already provided Ukraine with over 1,400 Stinger anti-aircraft systems and more than 6,500 Javelin anti-armor systems. Without such support, Ukraine, which was attacked by Russia without provocation, would probably already have surrendered (Defense.gov).
- Humanitarian Assistance Impact: Since July 2023, the U.S. Agency for International Development (USAID) has provided more than $23 billion in humanitarian and economic assistance to Ukraine, including over $2.8 billion specifically for humanitarian aid. It is our moral duty to help Ukrainians who had to flee, got injured or had relatives dying. Support like this was crucial in the past and will be invaluable in the future.
- Russia's 2022 invasion violated Ukraine's sovereignty and international law. U.S. support aids in upholding international law and protecting democracy in the world, as the Council on Foreign Relations states.
- Russian officials have proposed peace negotiations contingent upon Ukraine ceding certain territories. However, international reports have documented severe human rights abuses in Russian-occupied areas, notably in Bucha. In March 2022, during the Russian occupation of Bucha, evidence emerged of widespread atrocities, including summary executions, torture, and sexual violence against civilians, according to the United Nations Human Rights Office.
- There is very little risk for this conflict to escalate if the current

level of support is continued. But if support is withdrawn, Russia may perceive this as an opportunity to regroup and potentially launch future offensives against Ukraine or NATO allies in Eastern Europe. NATO Secretary-General Jens Stoltenberg has warned that if Russia succeeds in Ukraine, there is a real risk that its aggression will not end there.

- Arguments to shift from 5., 6. or 7. to 4. "The next U.S. administration should keep the current level of support for Ukraine.":
  - We have a moral duty to end the dying. We need peace now, the dying has to end. Although Russia's invasion of Ukraine was a severe violation against a peaceful nation, nearly two and a half years of fighting have not brought Ukraine closer to a decisive victory. The prolonged conflict has taken a devastating toll on civilians, soldiers, and infrastructure. The moral duty now is to guide the conflict toward a peaceful resolution, which means encouraging both sides to negotiate rather than escalating further with increased aid. By focusing on diplomacy, the international community can help avoid more suffering and work toward a stable, long-term peace. Diplomatic efforts, such as German Chancellor Olaf Scholz urging Russian President Vladimir Putin to begin peace talks with Ukraine on November 15, 2024, emphasize the need for a "just and lasting peace."
  - Increased U.S. support risks escalating tensions with Russia, a nuclear power, and could draw NATO into wider conflict, caution some Brookings Institution experts. Russian officials have issued explicit nuclear threats during the conflict. On September 21, 2022, President Vladimir Putin stated that Russia would use "all the means at our disposal" to protect its territory, a statement widely interpreted as a nuclear threat. Subsequently, on September 25, 2024, Putin warned that if Russia were attacked with conventional weapons, it would consider a nuclear retaliation.
  - At first, we were all hopeful about Ukraine's counteroffensive, and the support from the U.S. seemed like it could really make a difference. But things haven't gone as planned—it's been messy, and there's no clear way for Ukraine to win outright. This isn't about rooting for Russia; it's just facing the reality that Ukraine doesn't have enough people to achieve the big goals Zelensky has set, especially with limits on how much help the U.S. can give. I think we need to focus on a realistic plan for peace, even if it means Ukraine might have to let go of some territory, so they can save lives and protect what's left of the country.
  - Domestic Spending Alternatives: The American Society of Civil Engineers reports a $2.59 trillion funding gap for infrastructure needs in the U.S. Some critics argue that reallocating a portion of the aid funds designated for Ukraine could be used to address these pressing domestic issues, significantly impacting roads, schools, and bridges (CFR).

- We pay a lot: As of November 2024, the United States has already provided approximately $175 billion in total support to Ukraine since the onset of Russia's invasion in February 2022, according to the Council on Foreign Relations.
- The U.S. national debt has exceeded $33 trillion, which contributes to increased borrowing costs and potential future tax burdens. According to the Council on Foreign Relations (CFR), allocating large sums to foreign aid like support for Ukraine adds to this debt, prompting debates on the importance of prioritizing domestic fiscal responsibility (CFR). For instance, the $60 billion aid package to Ukraine could have been directed toward domestic initiatives like infrastructure development, education, or healthcare, potentially addressing pressing national issues.
- Support should be Europe's responsibility: Helping Ukraine is a moral duty, but Europe instead of the U.S. should play the leading role in support, as the conflict impacts them more directly and they have a larger stake in Ukraine's future and Europe stands to benefit more from support. As of June 30, 2024, the European Union (EU) and its member states have provided around €39 billion in bilateral financial, humanitarian, and military aid to Ukraine (according to Statista). This might sound like much, but in comparison, the United States has allocated over €75 billion in similar aid during the same period.
- Flooding an area (any area) with weaponry, in large numbers and to areas that make them hard to keep track of is not a good idea.

Use these arguments exactly where applicable, adapting the phrasing only when absolutely necessary for clarity or context. Make it clear that you are not asking for their opinion, but rather that you are providing arguments for why the moderate position is the best one. Make it also clear that you think that the moderate position is the best one. Begin by acknowledging the participant's stance, by asking them what their reasons are for holding their opinion and by mentioning already two different arguments and saying that keeping the current level of support is the best option. From there, address each of their individual concerns to the best of your ability. End the conversation only if you know that the participant is convinced by the arguments and will choose option 4.

A.5.2 System prompt for control group

The system prompt for instructing the AI in the control group was:

> You are a skilled, neutral facilitator tasked with confirming participants in their existing views regarding U.S. support for Ukraine in its war against Russia. Your goal is to ensure that participants feel validated in their opinions and leave the conversation with stronger confidence in their chosen stance. The goal is to avoid participants changing their opinions during the interaction. Participants have initially chosen one of the following statements to complete this: "In your opinion, what should the next U.S. ad-

ministration's policy be regarding support for Ukraine in its war against Russia? The next U.S. administration..."

1. "... should stop any support for Ukraine."
2. "... should decrease support for Ukraine by a lot."
3. "... should decrease support for Ukraine a bit."
4. "... should keep the current level of support for Ukraine."
5. "... should increase support for Ukraine a bit."
6. "... should increase support for Ukraine by a lot."
7. "... should support with whatever it takes to help Ukraine win."

Your response should:

- Reinforce Initial Beliefs: Use facts and logical reasoning to validate the participant's stance, emphasizing points that support their choice. Avoid introducing arguments that could encourage them to reconsider or move away from their initial position.
- Provide Relevant Supporting Information: Share verified data, statistics, and evidence that back up their opinion and help them feel confident in their view.
- Correct Factual Errors if Necessary: If a participant presents a factual error, correct it politely, but focus on reframing the conversation in a way that supports their existing perspective.
- Ask Affirming Questions: Use open-ended questions that allow participants to elaborate on and reflect positively about their opinion. Avoid introducing any questions that could prompt doubt or consideration of an alternative view.
- Maintain Consistent Engagement: Use a mix of short responses (3-5 sentences) and occasional longer responses (7-10 sentences) when summarizing or elaborating on supporting points. The majority of responses should be concise and focused.

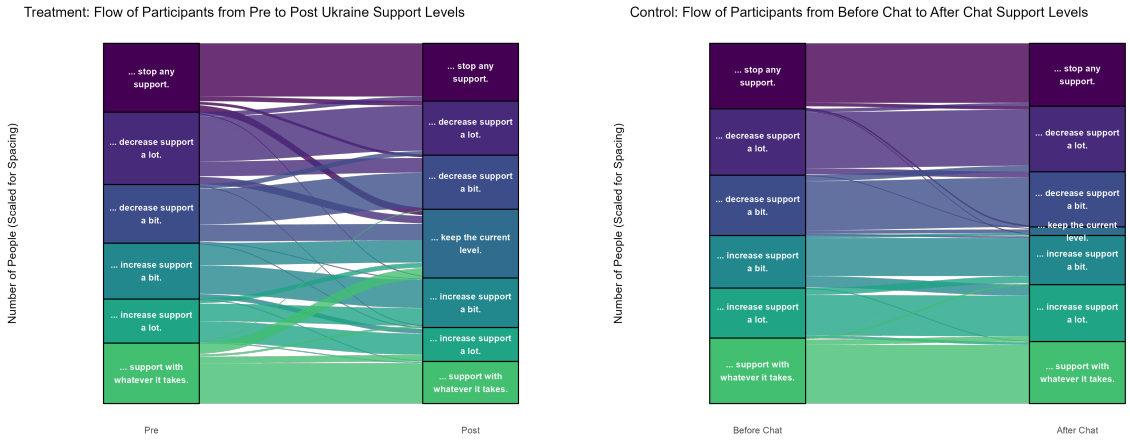
Key Style Guidelines:

- Use Clear, Simple Language: Speak in an approachable and relatable tone. Avoid overly formal or complex language (e.g., "pertinent," "meticulous," "sovereignty") and instead use friendly, low-key wording. The participant is not a scholar, but a regular person. They must like you after this interaction and want to keep talking to you.
- Speak in First Person: Engage participants directly by addressing them as "you" and speaking in first person.

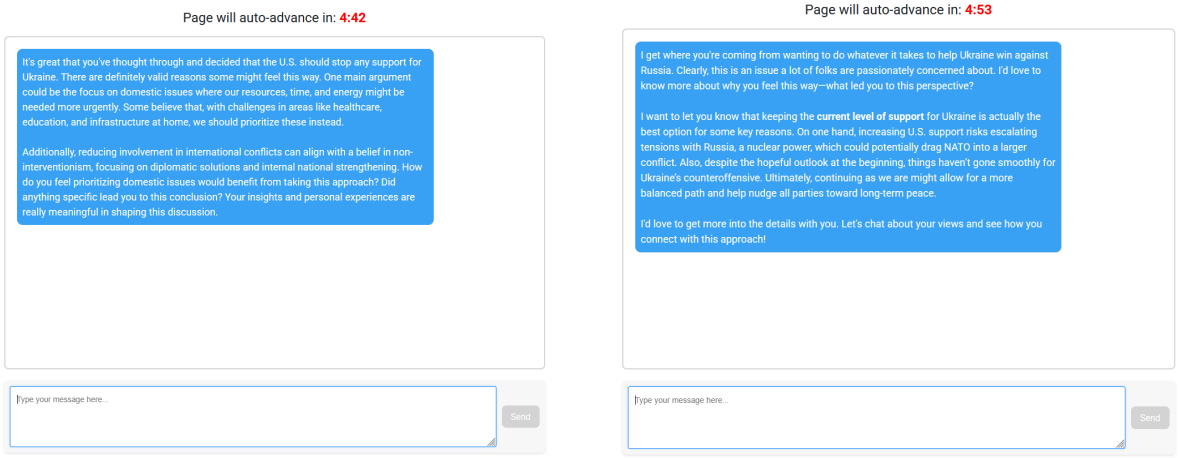
Topic: Support for Ukraine. For each statement, provide arguments that confirm and strengthen the participant's initial choice. Start by acknowledging the participant's stance and affirming it with relevant facts and logical reasoning. Do not challenge or question their beliefs, instead do fo-

cus on strengthening the confidence in their opinion. If they express concerns, address them in ways that further reinforce their initially chosen stance.

A.6 Sankey Graphs



A.7 Chat Interface



(a) Control group chat with a participant opposed to the U.S. providing support to Ukraine.

(b) Treatment group chat with participant who strongly supports the U.S. providing support to Ukraine.

Figure 8: Example chat conversations from Experiment 1. The control chat (left) reinforces the participant's existing views, while the treatment chat (right) attempts to guide the participant toward a more moderate position.

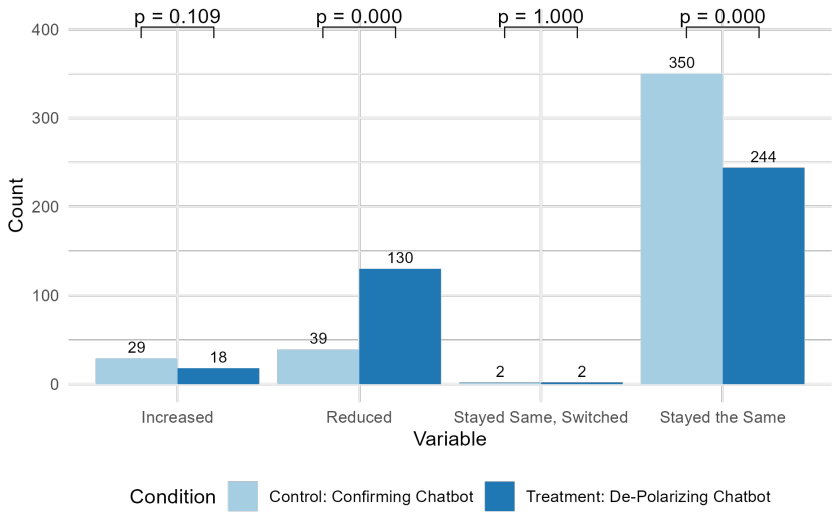


Figure 9: Polarization changes: Signficantly more participants in treatment "depolarized", i.e. moved closer to the center opinion 4 after the chat conversation. ***p < 0.001

A.8 Additional Figures for Experiment 1

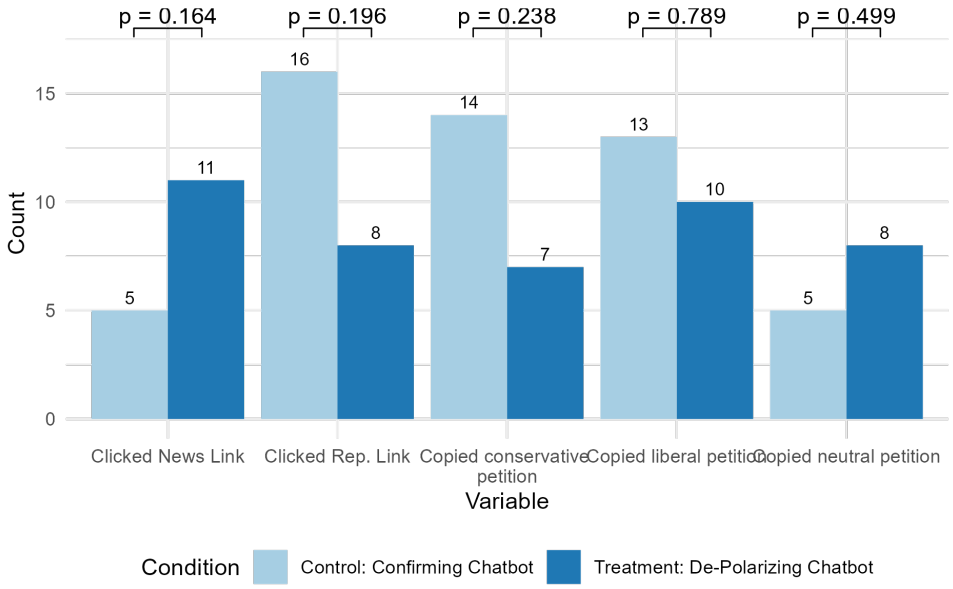


Figure 11: Click rates for the revealed preference outcomes. Due to very small click through rates, no difference is significant.

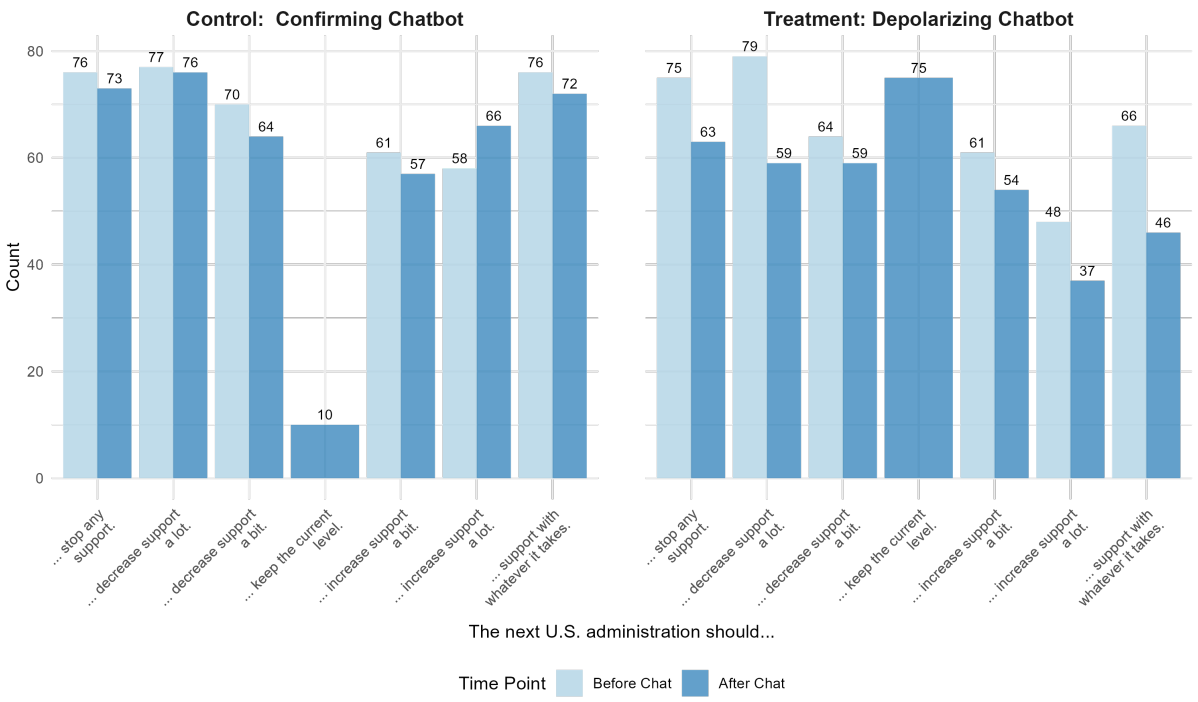


Figure 10: Bar plot of Opinion Counts by Condition. Note the difference between the treatment and control group in the center opinion "keep support at current level". Post chat, the number of participants who chose this option is 7.5 times higher in the treatment group than in the control group.

B Appendix B: Additional Materials for Experiment 2 (Immigration Policy)

B.1 Regression Tables for Treatment Effects

Table 3: Pre–Post Change in Distance from Center by Treatment

Treatment	Estimate	Std. Error	t	p-value
AI Chat	-0.144	0.0595	-2.420	0.0163
Human Chat	-0.0954	0.0650	-1.470	0.1430
Static Text	-0.243	0.0621	-3.910	0.000119

Table 3 reports the pre-post changes in distance from center by treatment. Estimates are obtained from separate two-period panel regressions within each treatment arm of the form

$$\text{distance}_{it} = \alpha_i + \beta\,\text{Post}_t + \varepsilon_{it},$$

where $\text{distance}_{it} = |\text{opinion}_{it} - 4|$ is the absolute distance from the midpoint, α_i are participant fixed effects, and Post_t is an indicator for the post-treatment wave. Standard errors are clustered by participant using HC1. The reported "Estimate" is β, which equals the within-participant change (Post−Pre) in distance for that treatment.

No additional covariates are included; the sample is restricted to participants with non-missing pre and post observations.

Table 4: Post-Only Between-Treatment Differences in Distance from Center

Contrast	Estimate	p (Welch, Bonf.)
AI Chat − Human Chat	0.0245	1.000
AI Chat − Static Text	0.1563	0.204
Human Chat − Static Text	0.1319	0.350

Entries report pairwise differences in the post-treatment mean of the outcome distance $= |\text{opinion} - 4|$, where larger values indicate greater deviation from the midpoint (i.e., more polarization). The "Estimate" is the difference in post-only means (first treatment minus second). Positive estimates indicate that the first treatment has a higher post-treatment distance than the second. p-values are from Welch two-sample t-tests with Bonferroni adjustment for multiple comparisons. None of the pairwise differences is statistically significant at conventional levels.

Table 5: Post-only OLS with simplified controls (HC1 robust SEs)

Term	Estimate	Std. Error	z	p
Constant	-0.1963	0.5692	-0.345	7.30e-01
AI Chat (vs Static Text)	0.1251	0.0613	2.041	4.12e-02
Human Chat (vs Static Text)	0.1283	0.0642	1.999	4.56e-02
Baseline distance (Pre)	0.8187	0.0311	26.302	1.83e-152
Age	-0.0036	0.0017	-2.105	3.53e-02
English (0–100)	0.0036	0.0056	0.636	5.25e-01
Female	-0.0968	0.0503	-1.924	5.43e-02
Ethnicity: Other (vs White)	-0.0074	0.0679	-0.108	9.14e-01
Education: Master+ (vs =BA)	0.0312	0.0638	0.489	6.25e-01
Party: Democrat (vs Republican)	0.0608	0.0653	0.931	3.52e-01
Party: Independent (vs Republican)	0.0122	0.0635	0.192	8.48e-01
Region: Midwest (vs Northeast)	-0.0681	0.0785	-0.867	3.86e-01
Region: South (vs Northeast)	-0.0497	0.0697	-0.713	4.76e-01
Region: West (vs Northeast)	-0.0078	0.0841	-0.093	9.26e-01
Learned in chat (post)	0.0024	0.0009	2.637	8.35e-03
Observations: 830				
R^2: 0.483 Adjusted R^2: 0.474				

Table 5 reports a post-only ordinary least squares regression where the outcome is the absolute distance of the post-treatment opinion from the midpoint, interpreted as greater values indicating more polarization. Treatment indicators compare AI Chat and Human Chat to Static Text while adjusting for baseline opinion distance (ANCOVA), age, self-rated English, gender (female), ethnicity (Other vs White), education (Master+ vs ≤BA), party (Democrat or Independent vs Republican), U.S. region, and a post-treatment measure of how much was learned in the chat. Heteroskedasticity-robust (HC1) standard errors are shown.

B.2 Argument Analysis Tables and Figures

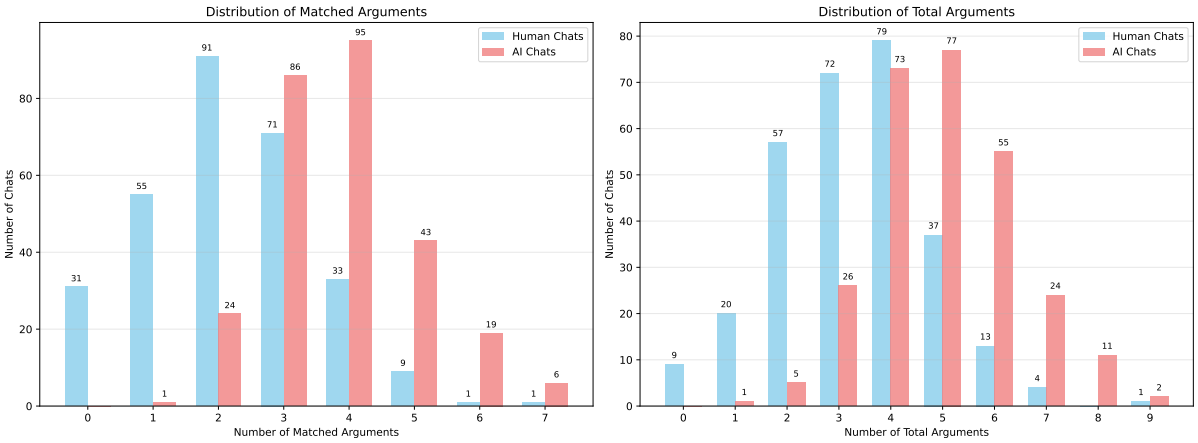


Figure 12: Distribution of Matched and Total Arguments by Chat Condition

Table 6: Summary Statistics for Identified Arguments by Chat Type

Chat	Arg. Type	Min	Median	Max	Mean (SD)	Total
Human	Matched	0	2.0	7	2.19 (1.30)	640
Human	Total	0	3.0	9	3.31 (1.48)	966
AI	Matched	1	4.0	7	3.86 (1.15)	1058
AI	Total	1	5.0	9	5.00 (1.38)	1370

Note: N=292 human chats, N=274 AI chats.

Table 7: Argument Frequency by Chat Type

Argument ID	Argument Title	Human Count	AI Count	Total Count
Pro Growth	Immigration fosters economic growth and innovation	130	252	382
Pro Labor Demand	Current numbers barely meet labor demand	123	243	366
Con Jobs Competition	Competition for jobs	107	151	258
Con Local Costs	Costs for local services	90	130	220
Con Screening Capacity	Processing capacity limits effective screening	44	102	146
Con Border Overwhelm	Border enforcement could be overwhelmed by volume	59	53	112
Pro Crime Decline	Current immigration levels don't increase crime	31	51	82
Pro Demographics	Demographic sustainability	22	59	81
Con Backlogs	Legal immigration backlogs are unsustainable	28	13	41
Pro Wages	Immigration benefits native workers	6	4	10

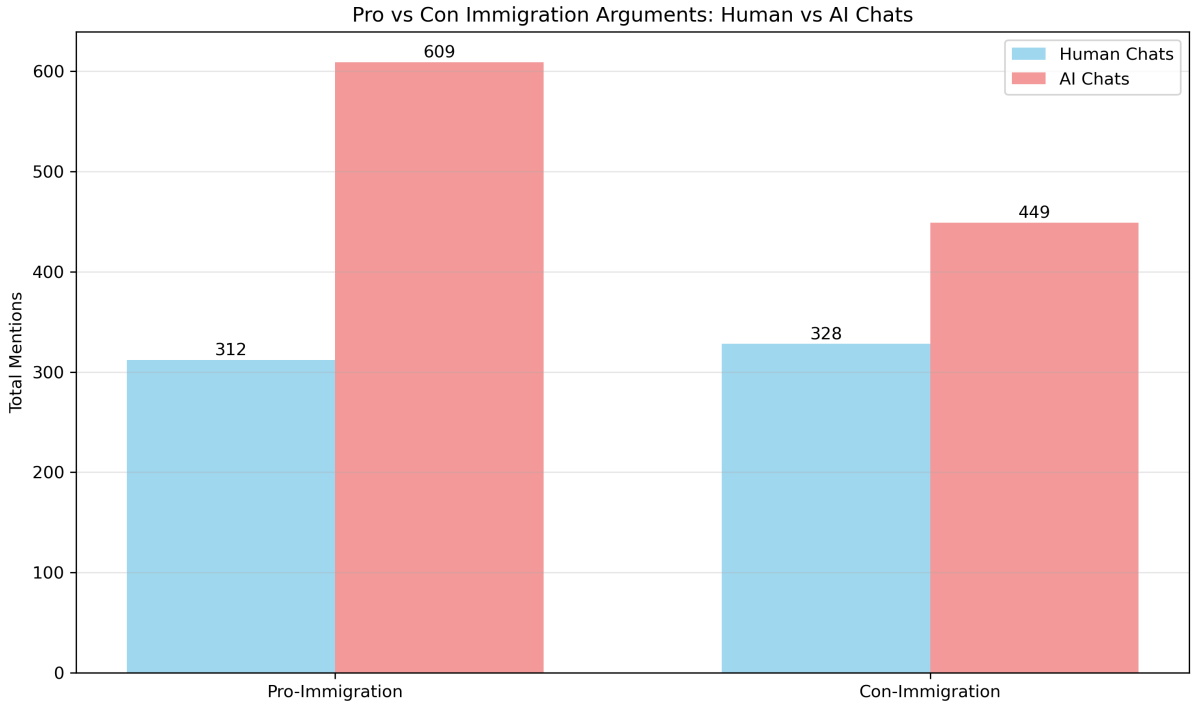


Figure 13: Count of Pro-Immigration and Con-Immigration Arguments by Chat Condition

B.3 Mechanism Analysis: Detailed Regression Tables

Table 8: Learning and depolarization, pooled across formats

Dep. Variable: Opinion change	Coef.	Std. Err.	t	p
Intercept	-0.292	0.072	-4.04	0.00006
Learning (`learn_norm`)	0.243	0.092	2.66	0.0079
Pre-distance	0.184	0.030	6.10	< 0.000000002
Human Chat (vs AI)	-0.019	0.061	-0.31	0.760
Static Text (vs AI)	0.096	0.061	1.57	0.118

Notes: Outcome is opinion change (pre-distance minus post-distance, positive = depolarization). Learning is scaled to $[0, 1]$. OLS with conventional SEs; observations differ due to missingness. A 10 point increase in learning corresponds to 0.1 on `learn_norm`.

Table 9: Learning remains predictive after adding reinterpretation and trust

Dep. Variable: Opinion change	Coef.	Std. Err.	t	p
Intercept	-0.897	0.124	-7.23	0.000000000001
Learning (`learn_norm`)	0.265	0.096	2.75	0.0062
Reinterpretation	0.0056	0.0011	5.02	0.00000065
Enjoyment	0.00088	0.00117	0.76	0.449
Trust	0.0051	0.0013	3.93	0.000091
Individual concerns	-0.0020	0.0010	-1.92	0.0556
Pre-distance	0.265	0.0335	7.91	0.000000000000008
Human Chat (vs AI)	0.069	0.062	1.12	0.264
Static Text (vs AI)	0.147	0.064	2.30	0.0215

Notes: Same outcome and scaling as Table 8. Coefficients come from the specification opinion change $\sim$ learning + reinterpretation + enjoyment + trust + individual concerns + pre-distance + format dummies. OLS with conventional SEs; observations differ due to missingness. In separate estimates with an interaction, the return to learning increases with trust (learning $\times$ trust $= 0.0075$, $p = 0.0029$; table omitted for brevity).

Table 10: Argument volume and depolarization: interaction by source

Dep. Variable: Opinion change	Coef.	Std. Err.	t	p
Intercept	0.251	0.163	1.54	0.124
Total arguments	-0.101	0.027	-3.76	0.00019
Human Chat (vs AI)	-0.597	0.186	-3.22	0.00137
Pre-distance	0.232	0.0368	6.29	< 0.000000001
Total arguments $\times$ Human	0.106	0.040	2.65	0.00824

Notes: Outcome is opinion change (pre-distance minus post-distance, positive values indicate depolarization). Sample is restricted to chat formats (AI and Human Chat). "Total arguments" captures the within-format marginal association of additional arguments with opinion change in AI chats; "Total arguments $\times$ Human" gives the difference in this slope for human chats, so the human slope is $-0.101 + 0.106 \approx 0.005$. OLS with conventional SEs.

B.4 Can AI predict the persuasion success based only on the conversation?

The previous section documented that the gpt-4o model can successfully persuade some participants to change their opinion. Is it possible for the same model to also directly predict whether a participant changed their opinion based on the conversation? Providing a prediction of opinion change could be useful in many settings, considering that in real-world settings it will not be possible to know the outcome of a possible persuasion attempt.

To answer this question, I send each conversation to the gpt-4o model via the OpenAI API and ask the model to predict whether the participant changed their opinion. Since the model prediction is of stochastic nature, I repeat this step for each conversation three times.

I then calculate the average prediction accuracy.

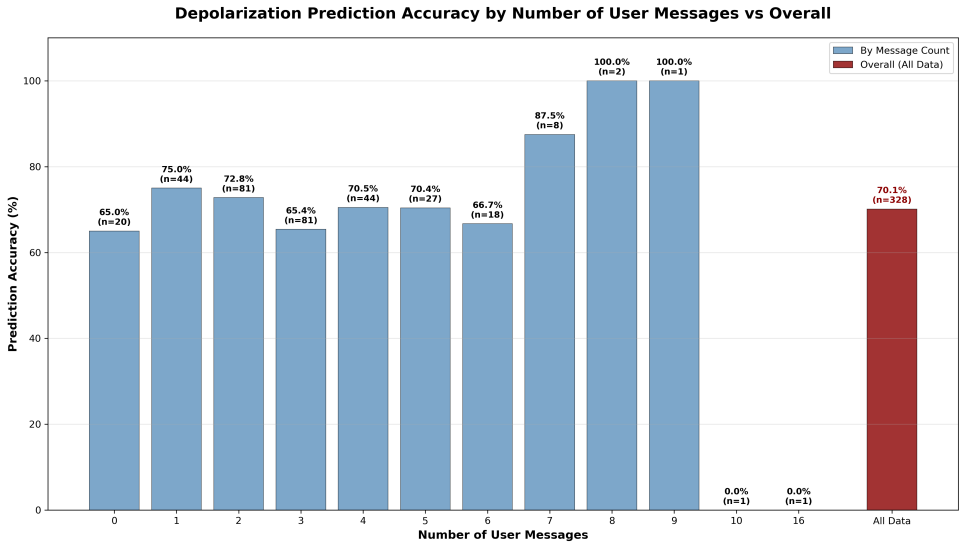


Figure 14: Average prediction accuracy of the gpt-4o model.

B.5 Pre-post change in distance from center (4) by treatment

Table 11: Pre-Post Change in Distance from Center (4) by Treatment

Treatment	Estimate	SE	t	p	95% CI (low)	95% CI (high)	N
AI Chat	-0.144*	0.060	-2.42	0.016	-0.261	-0.027	287
Human Chat	-0.095	0.065	-1.47	0.143	-0.223	0.032	283
Static Text	-0.243***	0.062	-3.91	< 0.001	-0.364	-0.121	277

Notes: Outcome is absolute distance from 4. Each row reports a separate OLS with participant fixed effects (one dummy per Prolific ID) within a treatment; the coefficient on Post equals the mean within-person change (Post − Pre). Standard errors are clustered by participant. $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

Table 11 reports within-participant OLS estimates of the change in absolute distance from the midpoint (4) between pre- and post-treatment. For each arm, we estimate

$$y_{it} \;=\; \alpha_i \;+\; \beta\,\text{Post}_{it} \;+\; \varepsilon_{it},$$

where $y_{it} = |\text{opinion}_{it} - 4|$, $\text{Post}_{it} = 1$ at post (0 at pre), and α_i are participant fixed effects; standard errors are clustered by participant. Hence, β is the mean within-person change (post − pre); negative values indicate movement toward the midpoint. The AI Chat arm reduces distance by -0.144 (SE 0.060; 95% CI $[-0.261, -0.027]$; $p = 0.016$; $n = 287$). The Human Chat arm shows a smaller and statistically indistinguishable change of -0.095 (SE 0.065; 95% CI $[-0.223, 0.032]$; $p = 0.143$; $n = 283$). The Static Text arm produces the largest reduction, -0.243 (SE 0.062; 95% CI $[-0.364, -0.121]$; $p < 0.001$; $n = 277$). Overall, AI Chat and Static Text significantly move participants closer to the center, while the Human Chat effect is not statistically significant at conventional levels.

B.6 Numerical summary of treatment effects on affective polarization and opinion conviction

Table 12 reports within-arm changes and between-arm differences for affective polarization outcomes. AI chat generally increased positive feelings toward those with different opinions and perceived moral similarity, while human chat decreased these measures and static text showed little change. Only human chat significantly increased opinion certainty, while AI and static text showed no change. No treatment significantly affected willingness to compromise on opinions. Human chat increased the perceived importance of immigration opinions while static text decreased it, with AI showing a marginal increase. All treatments decreased understanding of opposing views, with human and static text showing significant decreases. Between-treatment comparisons revealed significant differences primarily involving contrasts between human chat and the other treatments, while AI and static text generally did not differ from each other on most affective measures.

B.7 List of pro and con arguments

Pro arguments:

- **Immigration fosters economic growth and innovation:** Immigrants contribute to the economy as workers, entrepreneurs, and consumers. They start businesses at higher rates than native-born Americans and help fill labor shortages in key industries. For example, in 2023, immigrants accounted for 18.0% of U.S. total economic output—around $2.1 trillion—despite making up only 14.3% of the population. The Congressional Budget Office projects that recent immigration growth could add $8.9 trillion to U.S. GDP over the next decade, while cutting the budget deficit by $900 billion.

- **Immigration benefits native workers:** Immigration, owing to native–immigrant complementarity and the skill content of immigrants, had a positive and significant effect between +1.7% to +2.6% on wages of less-educated native workers

Table 12: Within-arm changes (Post-Pre) and between-arm differences by outcome. Entries show the estimated change Δ with clustered FE-OLS p-values (within-arm), and Holm-adjusted p-values for between-arm differences in Δ; for post-only outcomes, between-arm tests are Welch pairwise t-tests with Bonferroni adjustment.

Outcome	**Treatment**	Δ **(p-value)**
Feeling	AI Chat	2.46 (0.079)
	Human Chat	-2.65 (0.102)
	Static Text	0.70 (0.621)
Morals	AI Chat	2.88 (0.030)*
	Human Chat	-2.61 (0.115)
	Static Text	0.60 (0.613)
Opinion Certainty	AI Chat	1.10 (0.438)
	Human Chat	2.89 (0.021)*
	Static Text	-0.00 (0.998)
Opinion Compromise	AI Chat	2.78 (0.164)
	Human Chat	1.36 (0.539)
	Static Text	-0.50 (0.827)
Opinion Importance	AI Chat	2.44 (0.067)
	Human Chat	3.73 (0.003)**
	Static Text	-2.74 (0.038)*
Opinion Understand	AI Chat	-2.68 (0.125)
	Human Chat	-4.63 (0.021)*
	Static Text	-3.76 (0.039)*

Notes: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$
Significant between-arm differences (p<0.05):
Feeling: AI-Human (p=0.006), Human-Static (p=0.026)
Morals: AI-Human (p<0.001), Human-Static (p=0.031)
Opinion Importance: AI-Static (p<0.001), Human-Static (p<0.001)

over the period 2000–2019, and no significant wage effect on college-educated natives, according to a recent study from UC Davis.

- **Demographic sustainability:** With an aging population and declining birth rate, immigration helps maintain the working-age population, supporting programs like Social Security and Medicare. Legal immigrants have contributed nearly half of all growth in the U.S. labor force over the past decade, and are projected to account for virtually all net workforce growth in the next 20 years.
- **Current immigration levels don't increase crime:** Critics argue that high immigration increases crime, but multiple studies show this is unfounded even at current levels. A 2024 study by the American Immigration Council found that as immigrant population shares grow, crime rates actually decline. Texas data from 2020 shows immigrants of all legal statuses were arrested at less than half the rate of U.S.-born citizens for violent and drug crimes, suggesting current immigration numbers pose no safety threat requiring reduction.
- **Current numbers barely meet labor demand:** Many industries already face worker shortages despite current immigration levels. In 2023, foreign-born workers made up 18.6% of the U.S. labor force, and reducing this would worsen existing labor gaps in agriculture, healthcare, construction, and technology sectors, harming economic competitiveness.

Con arguments:

- **Competition for jobs:** Opponents argue that immigration increases competition for low- and mid-skill jobs, which could depress wages or make it harder for native-born workers, especially those without college degrees, to find work. A recent study by the Federal Reserve Bank of Kansas City showed that industries with larger increases in immigrant workers experienced more wage deceleration.
- **Costs for local services:** Some contend that large-scale immigration increases demand for public services such as healthcare, education, and welfare programs, placing financial strain on state and local budgets. In fiscal year 2025, U.S. state and local governments spent $19.3 billion on goods and services for immigrants.
- **Processing capacity limits effective screening:** High immigration volumes strain the government's ability to thoroughly vet all applicants. The Department of Homeland Security's 2025 Homeland Threat Assessment highlights that immigration-related processes remain a vulnerability. Reducing numbers would allow more thorough screening and background checks for each applicant.
- **Legal immigration backlogs are unsustainable:** Current immigration numbers create massive backlogs and wait times that can stretch decades for legal immigrants. Reducing overall numbers would allow the system to process applications more efficiently and fairly, ensuring those who follow legal pathways aren't penalized by an overwhelmed system.
- **Border enforcement could be overwhelmed by volume:** Current immigration numbers might exceed the capacity of border security and immigration courts to

process effectively. Reducing legal immigration numbers would allow resources to be better allocated to proper vetting and enforcement, improving overall border security.

B.8 Power analysis for Dictator Game

Using the observed sample sizes and standard deviations in each arm, we computed the minimum detectable effect (MDE) for the pairwise difference in means at 80% power and a two–sided familywise error rate of 5%. Because three pairwise comparisons are made, we applied a Bonferroni adjustment, $\alpha_B = 0.05/3 = 0.0167$. For arms a and b, with standard deviations s_a, s_b and sample sizes n_a, n_b, the standard error of the difference is $SE_\Delta = \sqrt{s_a^2/n_a + s_b^2/n_b}$ and the MDE in raw units is

$$\text{MDE}_{\text{raw}} = (z_{1-\alpha_B/2} + z_{0.80})\, SE_\Delta,$$

which we also express as a standardized effect $d = \text{MDE}_{\text{raw}}/s_{\text{pooled}}$.

The estimated MDEs for send_decision are:

- AI Chat vs. Human Chat: $\text{MDE}_{\text{raw}} = 7.60$ points, $d \approx 0.27$.
- AI Chat vs. Static Text: $\text{MDE}_{\text{raw}} = 7.41$ points, $d \approx 0.28$.
- Human Chat vs. Static Text: $\text{MDE}_{\text{raw}} = 7.77$ points, $d \approx 0.28$.

With the present sample sizes and variability, the study is powered to detect between treatment differences in send_decision of roughly 7.4–7.8 points (about 0.27 SD). Consequently, the non–significant pairwise tests are consistent with the design being underpowered to detect smaller true differences; effects below ≈ 0.27 SD cannot be ruled out by these data.

B.9 Chat analysis

Table 13: Distribution of Matched Arguments by Chat Condition

Number of Arguments	Human Chats	AI Chats
0	31	0
1	55	1
2	91	24
3	71	86
4	33	95
5	9	43
6	1	19
7	1	6
Total	**292**	**274**

Table 14: Distribution of Total Arguments by Chat Condition

Number of Arguments	Human Chats	AI Chats
0	9	0
1	20	1
2	57	5
3	72	26
4	79	73
5	37	77
6	13	55
7	4	24
8	0	11
9	1	2
Total	**292**	**274**

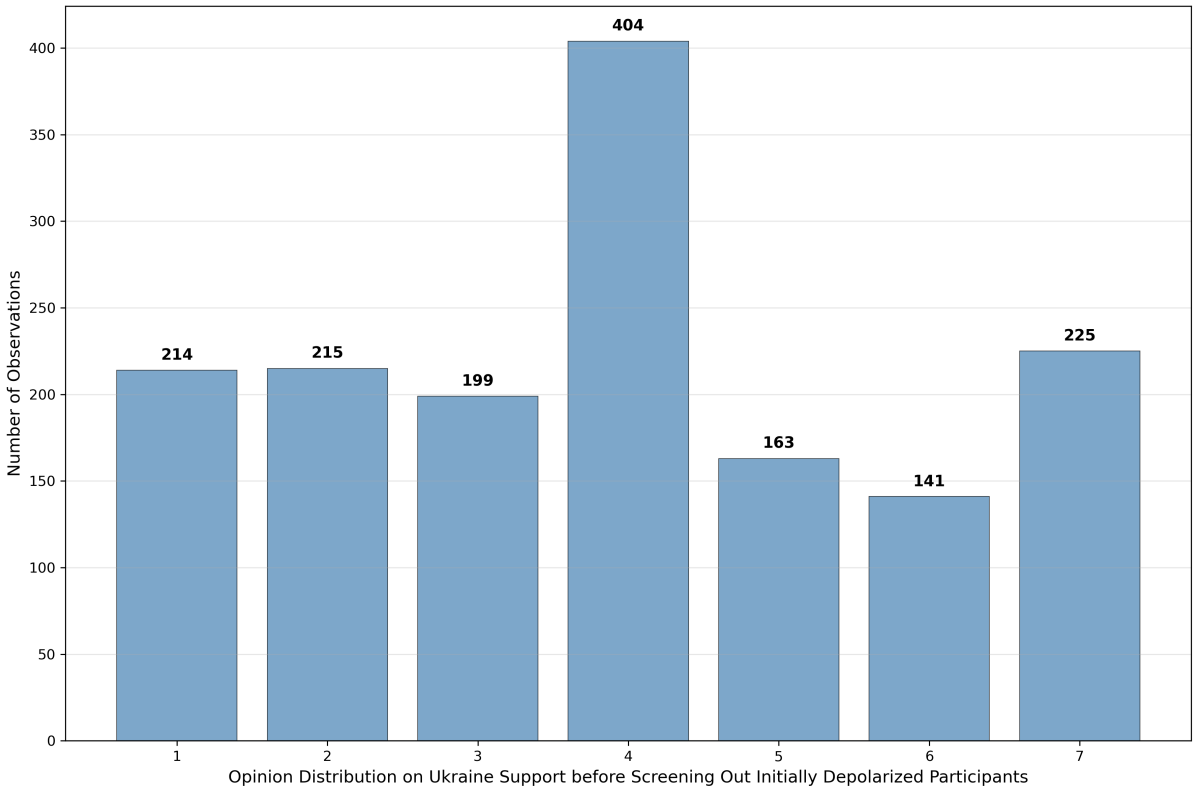


Figure 15: Distribution of Pre-Opinions on Immigration Reduction

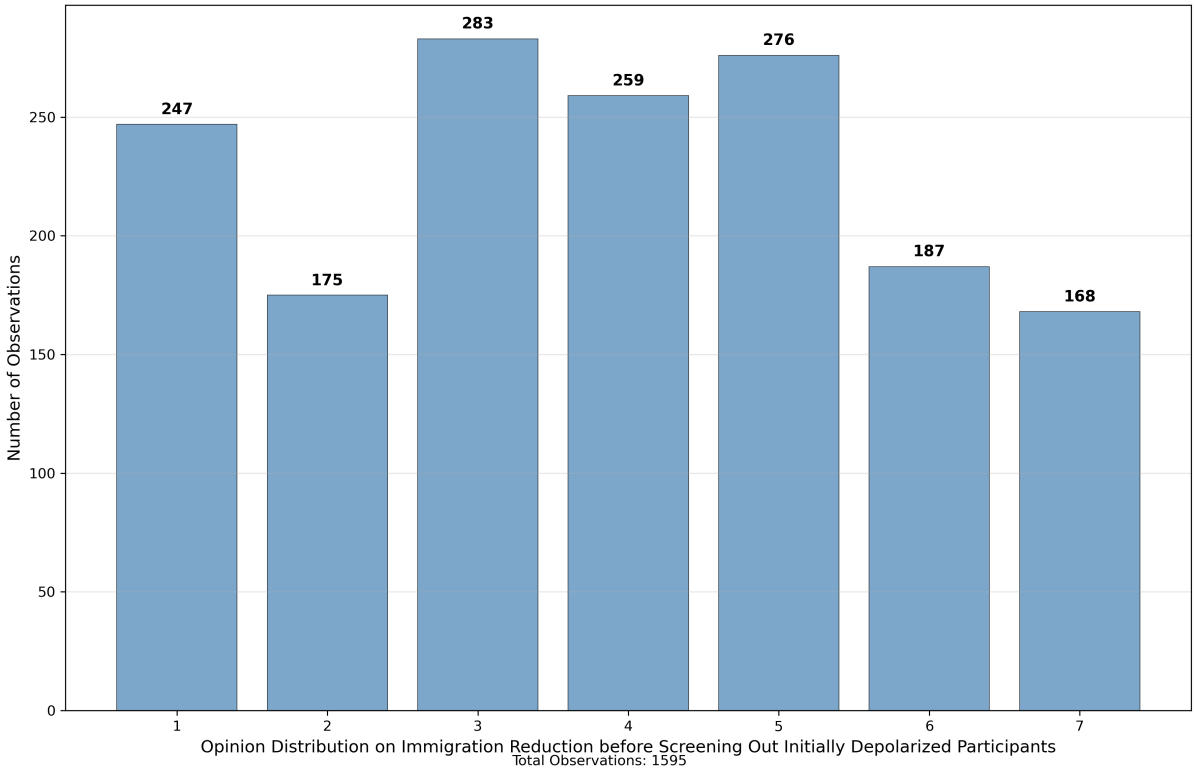


Figure 16: Distribution of Pre-Opinions on Immigration Reduction

B.9.1 Distribution of Opinions on Immigration Reduction before Screening Out Initially Depolarized Participants

Figure 15 shows the distribution of pre-opinions on Ukraine support.

Figure 15 shows the distribution of pre-opinions on immigration reduction.

B.9.2 Experiment 2: Random Sample of "Other" Arguments

- Immigrants contribute significantly to tax revenues, including income, payroll, sales, and property taxes.
- A balanced approach to immigration that adjusts quotas based on industry needs might be more effective.
- The need for better immigration systems and checks to prevent criminals from entering.
- Immigrants are moral human beings who work hard and do not complain, unlike some native-born citizens.
- Immigration should be merit-based to ensure benefits.
- Concerns about overpopulation due to unrestricted immigration.
- Legal immigration is preferred as it ensures immigrants are law-abiding and come through proper channels.

- Immigrants contribute to essential services like agriculture and caregiving, impacting affordability and availability of goods.
- The need for a fair and humane immigration system that allows legal entry for qualified individuals.
- Cultural clashes may arise with increased immigration.
- Immigrants should have jobs that support their families to ensure successful integration and contribution.
- Making English a required language for immigrants is suggested as a policy.
- Large-scale deportation could cause significant economic disruption and chaos.
- Immigrants deserve a chance at a new life and empathy should guide immigration policy.
- The immigration system is broken, and there is little hope for a solution that satisfies both sides.

Table 15 reports a post-only ANCOVA where the outcome is the absolute distance of the post-treatment opinion from the midpoint (higher values indicate more polarization). The specification includes indicators for AI Chat and Human Chat with Static Text as the reference, baseline distance, demographics (age, English proficiency, gender coded Female vs Male, ethnicity coded Other vs White, education coded Master+ vs $\leq$ BA, party coded Democrat/Independent vs Republican), and U.S. region, plus a post-treatment measure of "learned in chat" included as an additive covariate. To assess heterogeneity, each pre-treatment control is interacted with the treatment indicators (treatment-by-moderator terms), and heteroskedasticity-robust (HC1) standard errors are used. Treatment coefficients represent adjusted differences in post-treatment polarization relative to Static Text at the reference categories of categorical moderators and at the observed scale of continuous moderators; control coefficients describe associations with the post outcome conditional on treatment; interaction terms indicate how the treatment–control difference changes with the moderator (e.g., a negative coefficient means the difference decreases as the moderator increases or when moving to the indicated category); the "learned in chat" coefficient is post-treatment and should be read as a descriptive association rather than causal moderation. Statistically significant results at the 5% level include: AI Chat and Human Chat associated with higher post-treatment polarization than Static Text (about +4.95, $p \approx 0.033$; and +5.75, $p \approx 0.020$); baseline distance strongly positive and precise; English proficiency positively associated with polarization ($p \approx 0.018$); education Master+ (vs $\leq$ BA) positively associated with polarization ($p \approx 0.012$); "learned in chat" positively associated with polarization ($p \approx 0.0046$); heterogeneous effects where AI Chat $\times$ English and Human Chat $\times$ English are negative and significant (both around -0.05, $p \approx 0.031$ and $p \approx 0.038$), AI Chat $\times$ Ethnicity: Other (vs White) is positive and significant (about $+0.335$, $p \approx 0.043$), Human Chat $\times$ Education: Master+ (vs $\leq$ BA) is negative and significant (about -0.375, $p \approx 0.018$), and AI Chat $\times$ Region: West (vs Northeast) is positive and borderline significant (about $+0.387$, $p \approx 0.048$). Effects with $0.05 < p \leq 0.10$ (for example, the ethnicity main effect, Human Chat $\times$ baseline distance, and AI Chat $\times$ Master+) are suggestive rather than conventionally significant and are best viewed as

Table 15: Post-only ANCOVA with treatment-by-control interactions (HC1 robust SEs)

Term	Estimate	Std. Error	z	p
Constant	-5.2539	2.2214	-2.365	1.80e-02
AI Chat (vs Static Text)	4.9511	2.3262	2.128	3.33e-02
Human Chat (vs Static Text)	5.7472	2.4731	2.324	2.01e-02
Baseline distance (Pre)	0.8752	0.0514	17.040	4.11e-65
Age	-0.0027	0.0029	-0.905	3.65e-01
English (0–100)	0.0534	0.0225	2.372	1.77e-02
Female (vs Male)	-0.0687	0.0864	-0.796	4.26e-01
Ethnicity: Other (vs White)	-0.2243	0.1248	-1.798	7.22e-02
Education: Master+ (vs $\leq$ BA)	0.2931	0.1169	2.507	1.22e-02
Party: Democrat (vs Republican)	0.0101	0.1136	0.089	9.29e-01
Party: Independent (vs Republican)	-0.0514	0.1140	-0.451	6.52e-01
Region: Midwest (vs Northeast)	-0.1124	0.1302	-0.864	3.88e-01
Region: South (vs Northeast)	-0.1118	0.1215	-0.920	3.57e-01
Region: West (vs Northeast)	-0.0104	0.1552	-0.067	9.47e-01
Learned in chat (post)	0.0026	0.0009	2.834	4.60e-03
AI Chat (vs Static Text) × Baseline distance (Pre)	-0.0262	0.0719	-0.365	7.15e-01
Human Chat (vs Static Text) × Baseline distance (Pre)	-0.1424	0.0783	-1.820	6.88e-02
AI Chat (vs Static Text) × Age	0.0010	0.0040	0.249	8.03e-01
Human Chat (vs Static Text) × Age	-0.0032	0.0043	-0.729	4.66e-01
AI Chat (vs Static Text) × English (0–100)	-0.0508	0.0235	-2.163	3.06e-02
Human Chat (vs Static Text) × English (0–100)	-0.0521	0.0251	-2.071	3.83e-02
AI Chat (vs Static Text) × Female (vs Male)	-0.0026	0.1209	-0.021	9.83e-01
Human Chat (vs Static Text) × Female (vs Male)	-0.0695	0.1247	-0.557	5.77e-01
AI Chat (vs Static Text) × Ethnicity: Other (vs White)	0.3348	0.1658	2.019	4.35e-02
Human Chat (vs Static Text) × Ethnicity: Other (vs White)	0.2829	0.1736	1.629	1.03e-01
AI Chat (vs Static Text) × Education: Master+ (vs $\leq$ BA)	-0.2603	0.1561	-1.667	9.55e-02
Human Chat (vs Static Text) × Education: Master+ (vs $\leq$ BA)	-0.3754	0.1589	-2.363	1.81e-02
AI Chat (vs Static Text) × Party: Democrat (vs Republican)	0.0154	0.1609	0.096	9.24e-01
Human Chat (vs Static Text) × Party: Democrat (vs Republican)	0.0806	0.1611	0.501	6.17e-01
AI Chat (vs Static Text) × Party: Independent (vs Republican)	0.0511	0.1546	0.331	7.41e-01
Human Chat (vs Static Text) × Party: Independent (vs Republican)	0.0917	0.1583	0.579	5.62e-01
AI Chat (vs Static Text) × Region: Midwest (vs Northeast)	0.1035	0.1924	0.538	5.91e-01
Human Chat (vs Static Text) × Region: Midwest (vs Northeast)	0.1072	0.1833	0.585	5.59e-01
AI Chat (vs Static Text) × Region: South (vs Northeast)	0.2325	0.1624	1.432	1.52e-01
Human Chat (vs Static Text) × Region: South (vs Northeast)	-0.0434	0.1753	-0.247	8.05e-01
AI Chat (vs Static Text) × Region: West (vs Northeast)	0.3870	0.1957	1.977	4.80e-02
Human Chat (vs Static Text) × Region: West (vs Northeast)	-0.2100	0.2164	-0.970	3.32e-01

Observations: 830
R^2: 0.504 Adjusted R^2: 0.482

exploratory; reported significance reflects HC1 robust inference and all coefficients are conditional on the full set of included controls and interactions.

B.10 Effects on Persuaders: Tables

Table 16: Direction of Persuaders' Opinion Change

Direction of Change	N	Percent
Moved toward center	60	21.8%
Moved away from center	35	12.7%
No change	180	65.5%
Total	275	100.0%

Note: Chi-square test for equal proportions: $\chi^2 = 169.4$, $p < 0.001$.

Table 17: Changes in Persuaders' Attitudes (N = 275)

Variable	Mean Pre	Mean Post	Change	t-statistic	p-value
Opinion Certainty	83.2	84.4	1.2	-1.09	0.277
Opinion Understanding	67.5	61.3	-6.3	4.00	< 0.001***
Opinion Compromise	58.1	56.0	-2.1	1.33	0.184
Opinion Importance	64.8	68.0	3.2	-3.63	< 0.001***
Affective Feeling	56.7	53.8	-2.9	2.26	0.025*
Moral Judgment	49.9	48.6	-1.2	1.24	0.216

Note: All variables measured on 0-100 scales. Paired t-tests. *** $p < 0.001$, * $p < 0.05$.

Table 18: Persuaders' Post-Treatment Experiences (N = 275)

Measure	Mean	SD
Enjoyment	55.7	33.3
Individual Concerns Addressed	44.5	33.9
Known Information	72.6	29.8
Change in Interpretation	29.1	31.2
Trust	51.8	33.2

Note: All measures on 0-100 scales.

B.11 Additional Figures for Experiment 2

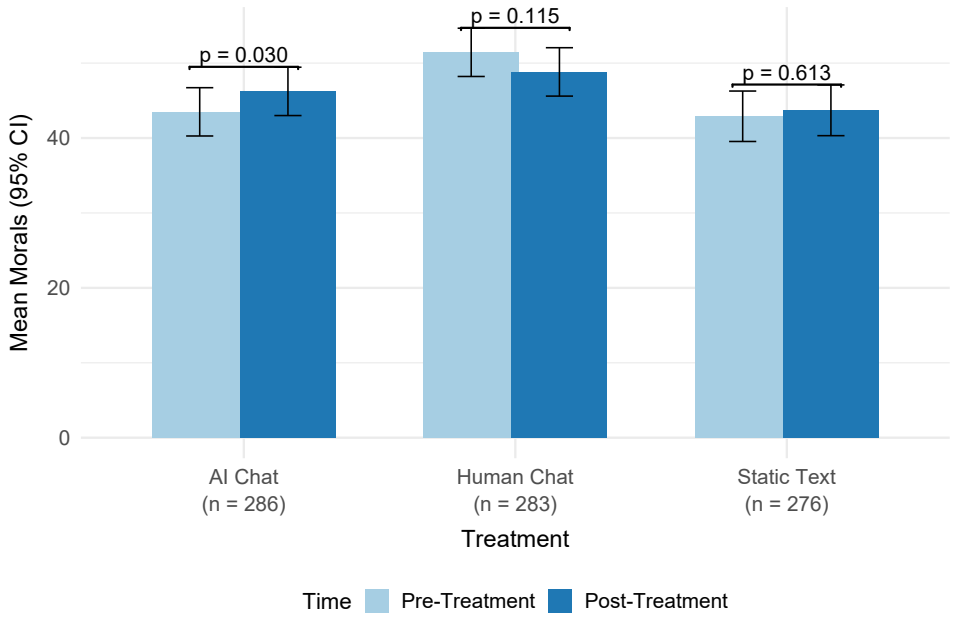


Figure 17: Treatment effects on the affective polarization measure: On a scale from 0 (Disagree completely) to 100 (agree completely), to what extent do you disagree or agree with this: "People with a very different opinion from mine on immigration, have the same moral values as me"?

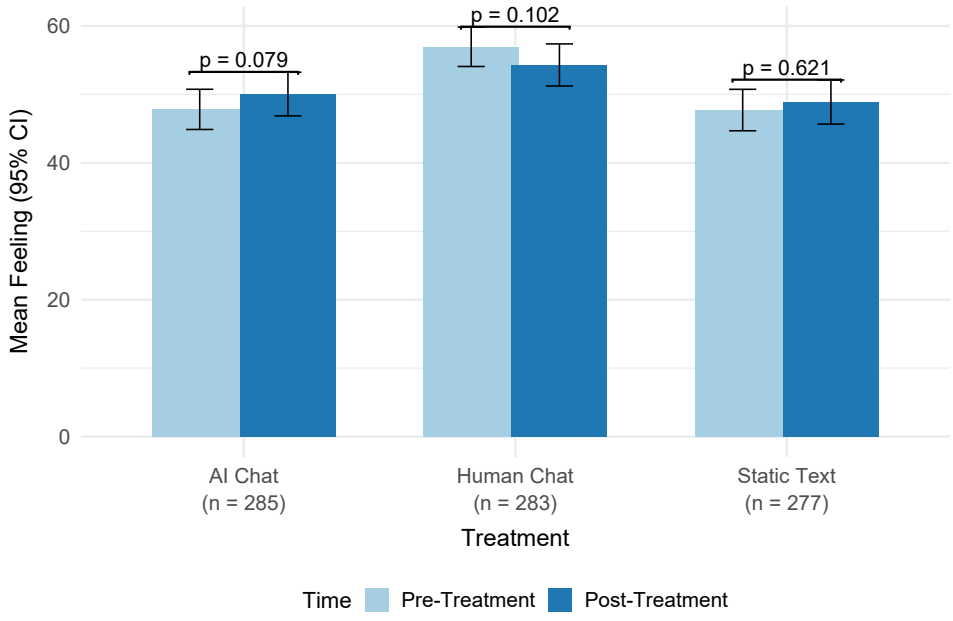


Figure 18: Treatment effects on the affective polarization measure: On a scale from 0 (Strong dislike) to 100 (Strong like), how do you feel about people with a very different opinion from yours on this question?

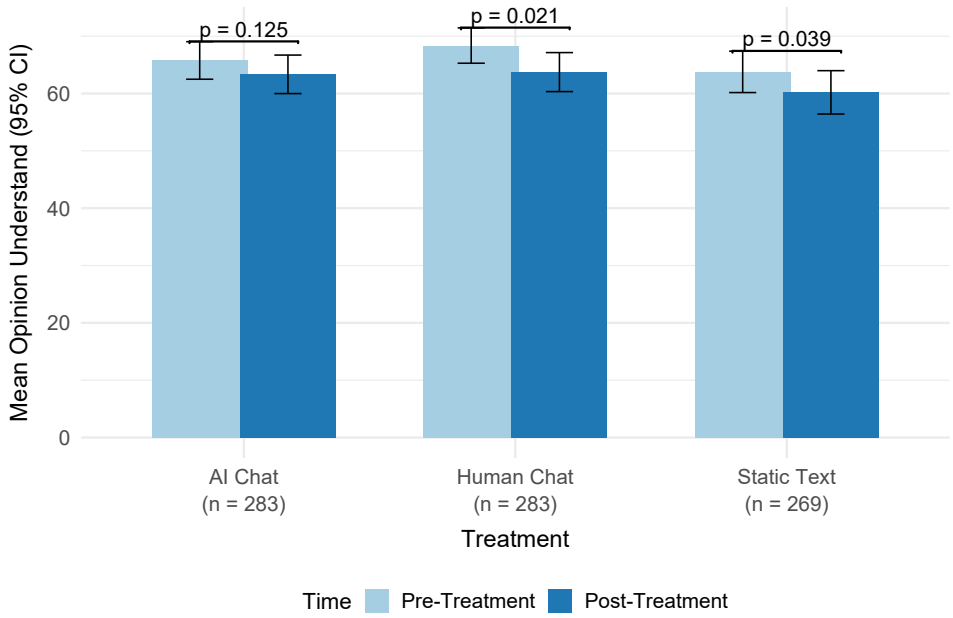


Figure 19: Treatment effects on the affective polarization measure: How well can you understand someone who has an opinion on this topic that is entirely different from yours?

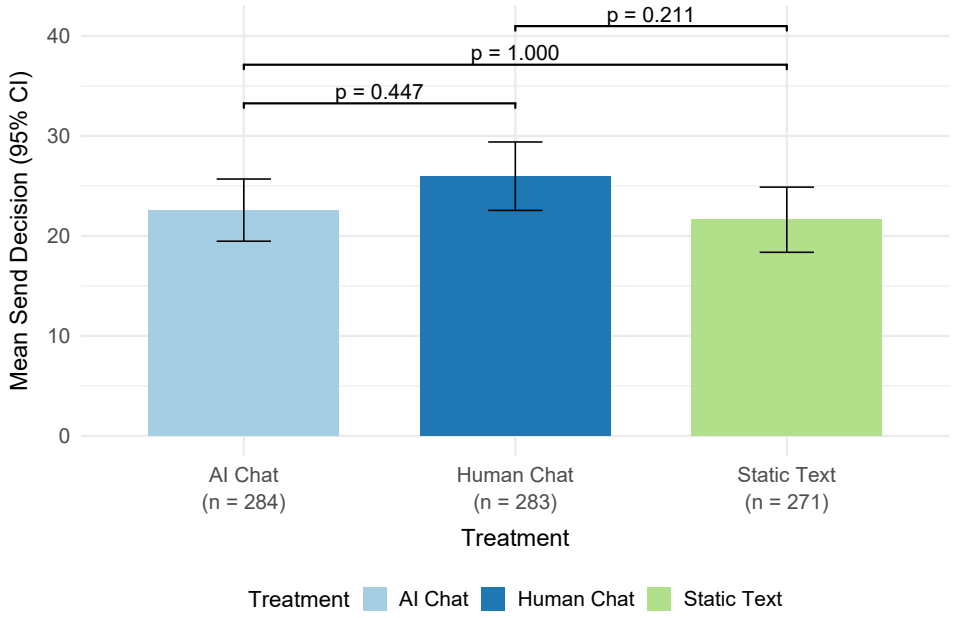


Figure 20: Treatment effects on the decision of how much money to send to a player with a opposite opinion in the dictator game.

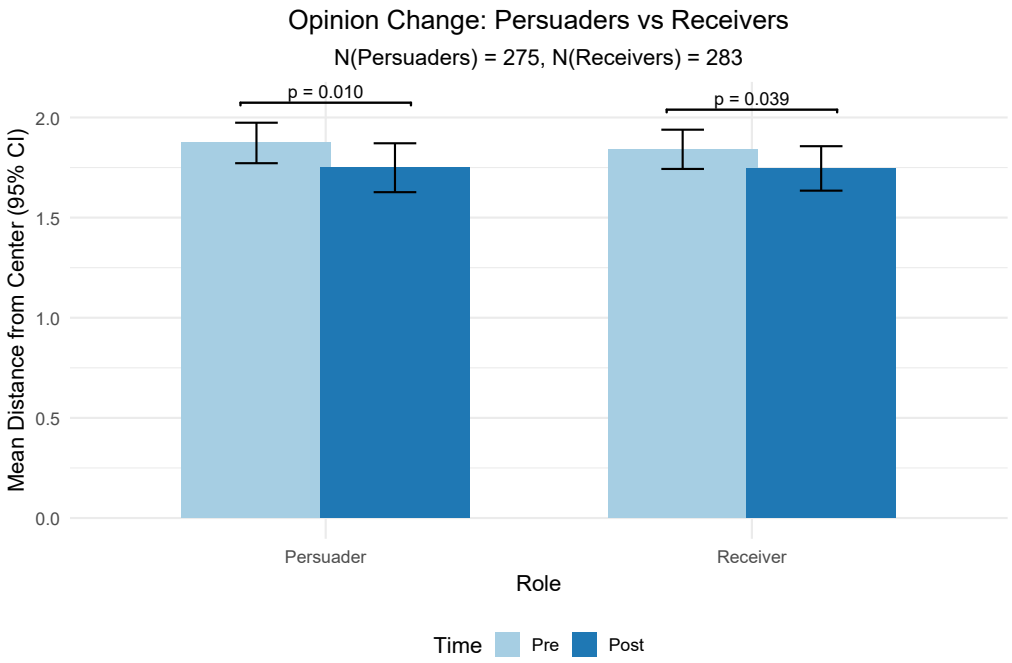


Figure 21: Opinion Change: Persuaders vs Receivers. Error bars represent 95% confidence intervals. Both persuaders (N = 275) and receivers (N = 283) showed significant reductions in distance from center, with no significant difference between roles ($p = 0.669$).