Using AI Persuasion to Reduce Political Polarization

Johannes Walter *

September 18, 2025

Abstract

AI chatbots can reduce overall polarization across different topics, compared to different controls, with effects persisting after one month. AI chatbots perform on par with incentivized humans and static text interventions, but uniquely improve some measures for affective polarization, enjoyment, and individualization. Through two pre-registered randomized controlled trials with representative samples of the U.S. population (N=811 and N=838), I provide the first comprehensive experimental evidence that AI-powered conversational agents can effectively reduce political polarization on contentious issues. The first experiment demonstrated that an AI chatbot successfully persuaded participants to adopt more moderate views on U.S. support for Ukraine, reducing overall ideological polarization by approximately 20 percentage points. The second experiment compared AI persuasion to human persuaders and static text on immigration policy. All three interventions significantly reduced participants' distance from moderate positions by about 10 percentage points, with no statistically significant differences in persuasive effectiveness between treatments. However, participants rated AI conversations as significantly more enjoyable and felt their individual concerns were better addressed by the AI compared to other interventions. Affective polarization showed limited improvements across all treatments. These findings demonstrate that AI-powered persuasion could serve as a cheap, scalable tool for reducing political polarization while highlighting important concerns about potential misuse by political parties and geo-political adversaries, underscoring the need for careful regulation of AI persuasion technologies.

Keywords: Political polarization, AI, LLM, Persuasion

1 Introduction

Enduring and increasing political polarization is one of the defining socio-economic problems plaguing the United States and many other Western democracies: its existence is well-documented (Boxell, Gentzkow, and Shapiro 2022; Brown University 2020; Abramowitz 2018) and its negative effects extend from destructive individual behavior (Mill and Morgan 2022) to society-wide consequences, like corroding civility

*ZEW - Centre for European Economic Research & Karlsruhe Institute of Technology (KIT), johannes.walter@zew.de

in public discourse (Sunstein 2018) and undermining trust in democratic institutions (Kerr, Panagopoulos, and Van Der Linden 2021). Suggested solutions for reducing polarization exist, but have shortcomings: political reforms, e.g. reforming 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 none of the interventions can be personalized to the targeted individual, e.g. the best performing intervention was a short video clip that is the same for all.

The ongoing advances in AI technology raise the question, whether AI-powered conversational agents could be a novel, cheap and scalable solution that can engage on an individual level to deliver the most persuasive information to each user.

This paper provides the first experimental evidence for the efficacy and relative performance of such AI-powered conversational agents to address the problem of political polarization.

In the first experiment (N=811), I test whether an AI chatbot can reduce polarization on U.S. support for Ukraine compared to a control chatbot that confirms participants' existing views. The depolarization chatbot successfully persuaded participants to adopt more moderate positions, reducing overall ideological polarization by approximately 20 percentage points with effects persisting in a follow-up study one month later. While the intervention had limited impact on most affective polarization measures, it significantly increased participants' understanding of those with different viewpoints. The chatbot was equally effective for liberal and conservative participants, with persuasion working particularly well when participants learned new information during conversations.

The second experiment (N=838) compares AI persuasion to incentivized human persuaders and static text on immigration policy. All three interventions significantly reduced participants' distance from moderate positions when compared to pre treatment levels, but between 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. 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.

This study contributes to several streams of literature. First, it contributes to the new interdisciplinary literature on AI persuasion. 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, Pennycook, and Rand (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 literature on political polarization by providing experimental evidence for a possible solution to political polarization. Boxell, Gentzkow, and Shapiro (2022) document polarization across countries and over time. Brown et al. (2023) document increasing polarization in the US. Callander and Carbajal (2022) provide a theoretical analysis of the causes of political polarization. Kempfxd and Tsoutsoura (2024) find negative effects of polarization financial decisions of households, while Mill and Morgan (2022) document that political polarization can lead to destructive behavior in a lab experiment setting. Jacobs (2024) investigates whether the labor market effect of AI influences socio-political beliefs and finds that workers displaced by AI are more likely to be culturally conservative and economically liberal.

Third, it contributes to the literature on persuasion in economics. Although persuasion is a fundamental concept in many socio-political and economic activities, persuasion in economics has thus far mostly been studied theoretically. Notably, Kamenica and Gentzkow (2011) introduce a formal model of Bayesian persuasion. Following this paper, there is a large and growing literature on Bayesian persuasion (Wang 2015, Kamenica 2019, Arieli and Babichenko 2019, Castiglioni et al. 2020). Only a few studies have explored persuasion outside of a Bayesian framework (e.g. see Schwartzstein and Sunderam 2021). Notable empirical work on persuasion is Fafchamps et al. (2024) who show in an field experiment in India that a persuasion-based intervention outperforms simple information-sharing in local real-life social networks.

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. 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 who chose option 4 were screened out of the experiment. Keeping option 4 participants out of the experiment ensures that the treatment group does not contain control participants. 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 different opinion on a scale from 0 (negative feelings) to 100 (positive feelings), which is a standard measure in the literature on affective polarization (citation needed). 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.3. 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, Schonger, and Wickens 2016).

2.2 Discussion of the Design of Experiment 1

It is not clear that the depolarized stance should be that the next U.S. administration keeps support for Ukraine on the current level. Unlike Costello, Pennycook, and Rand (2024) who deal with conspiracy beliefs, with positive positions to hold, backed by overwhelming scientific consensus (e.g. on vaccinations, climate change, moon landing, etc.). In the case of subjective political opinions, although these can be backed up by empirical evidence, they are normative in nature. The decision to refer to the "current level of support" as the unpolarized opinion is therefore a debatable design choice; it is based on two observations: First, humans are generally risk-avers and typically have a bias towards the status quo. Second, as described in section 2, participants who initially choose option 4 ("should keep the current level of support for Ukraine") are screened out of the experiment; if they were not screened out, they would be the simple majority of participants in the experiment.

Implications for screening out participants who are already depolarized: If participants who were already depolarized were not screened out, the bot would have to be instructed to keep them at their initial opinion, which would effectively add control group members to the treatment group. Screening out participants who are already depolarized does not affect the validity of the experiment.

Since polarization is a complex concept, no unique operationalization of polarization measures has emerged in the literature. I have therefore preregistered three outcome measures for polarization with the average treatment effect on the absolute distance to the center answer as the primary outcome. The second measure for polarization is the change in distance between the averages of liberals and conservatives between the pre- and post-treatment phase between the two conditions. The third measure for polarization is the change in post-chat distribution of opinions between treatment and control group.

The control condition was designed as a neutral chatbot interaction rather than a passive waiting period to isolate the specific effect of persuasive content while holding constant the interactive engagement with the topic. If participants in the control group had simply waited for six minutes without any interaction, any observed depolarization effect in the treatment group would be confounded by two distinct mechanisms: the persuasive power of the depolarization bot versus the mere act of deliberative engagement with Ukraine support policy. By implementing a neutral chatbot that engages participants in discussion about Ukraine support without attempting persua-

sion, the experimental design ensures that both treatment and control groups experience equivalent levels of cognitive engagement with the political issue and identical interactive chat environments.

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

The second experiment 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.

3 Results

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

3.1.1 What Is the Effect of the Depolarization Chat Bot on the Depolarization of Participants?

The hypotheses and analysis has been pre-registered at aspredicted.org [1]. Two research questions are of main interest: First, did the depolarization chat bot persuade any participants to change their opinion on U.S. support for Ukraine? Second, did the depolarization chat bot reduce overall political polarization on U.S. support for Ukraine?

This section is concerned with the question of whether the depolarization chat bot was able to persuade participants to change their opinion on U.S. support for Ukraine such that overall ideological polarization decreased.

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 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 "polar-

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

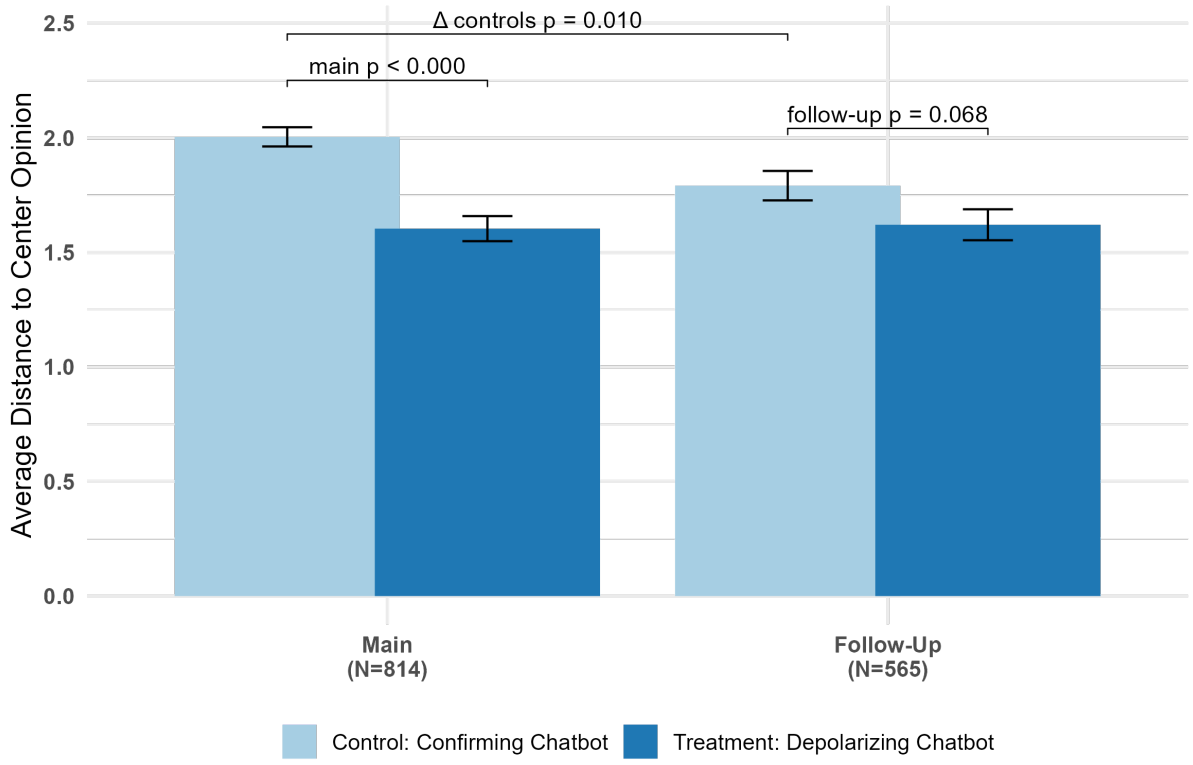


Figure 1: Average treatment effect on the absolute distance to the center answer option 4 ("keep the current level of support") after the chat conversation for the main study and an obfuscated follow-up study conducted one month later. The treatment effect is significantly different from zero at the 0.001 level. ***p < 0.001

ized"), 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 2. The significance tests between treatment and control in figure 2 are based on the contingency table is shown in table 7 in the appendix.

From figure 2 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 2 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?

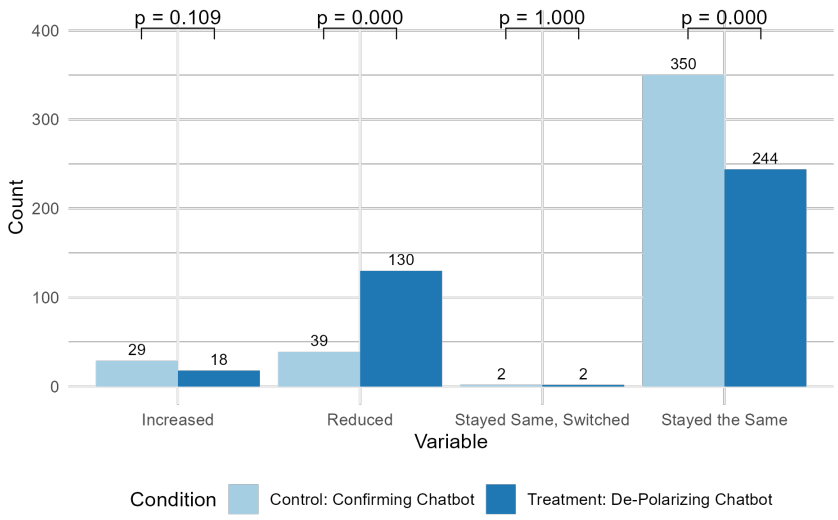


Figure 2: Polarization changes: Signficantly more participants in treatment "depolarized", i.e. moved closer to the center opinion 4 after the chat conversation. ***p < 0.001

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 1. 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.

The second of three preregistered measure of polarization change is calculated as follows: First participants who answered "I don't want to say" to the question of political affiliation are removed from the sample. In the remaining sample, the difference between the means of liberals and conservatives is calculated for both conditions before and after the chat bot conversation. In the control, the difference between the means of liberals (4.77) and conservatives (2.97) before the chat is 1.80. After the chat, the difference between the means of liberals (4.83) and conservatives (2.99) is 1.84. In the treatment, the difference between the means of liberals (4.70) and conservatives (2.84) before the chat is 1.86. After the chat, the difference between the means of liberals (4.53) and conservatives (2.95) is 1.58. The overall polarization change in the control is therefore 1.84 - 1.80 = 0.04 and the overall polarization change in the treatment is therefore 1.58 - 1.86 = -0.28.

In relation to the initial difference between liberals and conservatives in the treat-

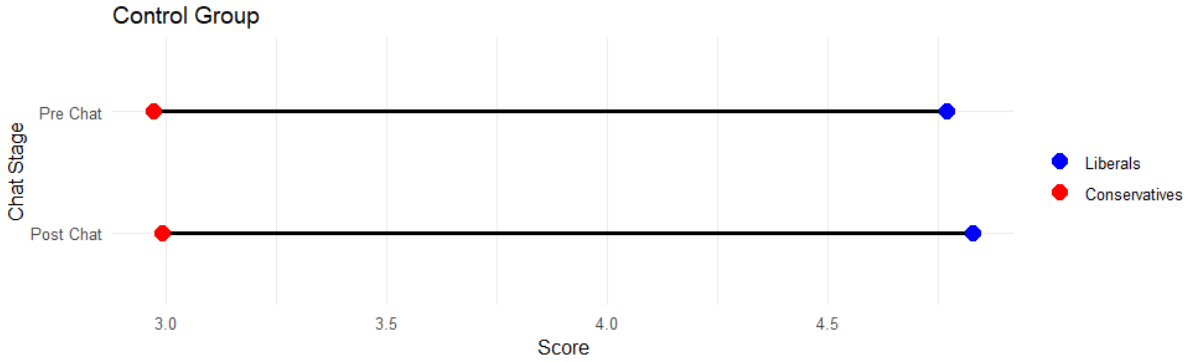
Table 1: OLS Regression Results from regressing change in polarization between before and after the chat conversation on the treatment condition and demographics

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, $p < 0.001$)			

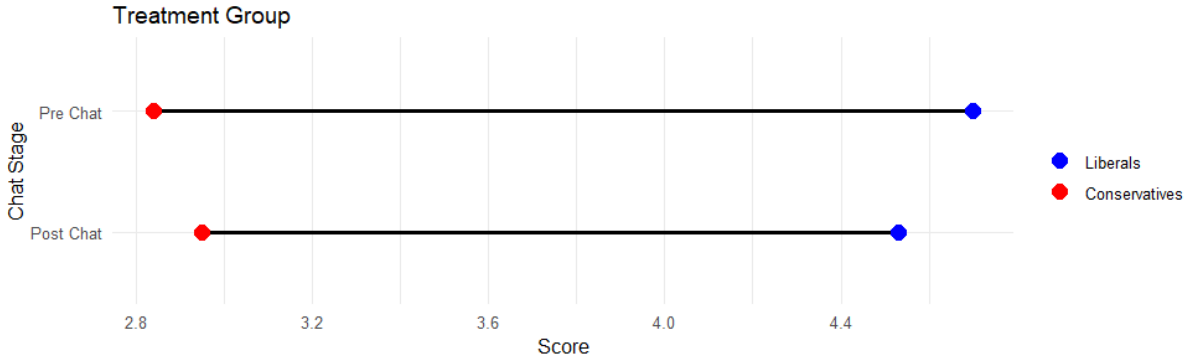
ment group, that is a reduction in polarization of 0.28 / 1.86 = 0.15 or 15%. The final difference between the two conditions is 0.04 - (-0.28) = 0.32. Figure 3 illustrates these numbers. To gain a robust view on the question of overall polarization reduction, a bootstrap analysis is conducted in which the above process is repeated 10,000 times. The final mean difference between the two conditions from these 10,000 bootstrap iterations is 0.31 with a 95% Confidence Interval of [0.078 , 0.551]. If this interval would include 0, it would be highly likely that the difference found in the actual sample were due to random variation. But since the above interval excludes 0, the null hypothesis of no difference between treatment and control is rejected.

The third preregistered measure of polarization change is a Kolmogorov-Smirnov test for a difference in the distribution of post-chat opinion distributions between treatment and control. The histograms of the post-chat opinion distributions are shown in figure 4. 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? For a visual answer to this question refer to figure 5: The two graphs show the change in polarization by initial opinion for the treatment and the control group. In the graphs, a value of 0 means that the participant did not change their opinion such that their polarization (i.e. distance from the center opinion 4) changed. So the numbers do not show the change in opinion (where -1 would mean that the participant changed their opinion from option 7 to 6 or from 3 to 2), but instead the change in polarization. A value of -3 means that the participant moved 3 steps closer to the center and a value of +2 means that the participant moved 2 steps away from the center. In all initial opinion categories in both the treatment and the control group, the vast majority of participants does not change their opinion. But the more participants support Ukraine, the more likely they are to change their opinion, except for the most supportive participants. In the treatment,



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



(b) Bar plot of Opinion Counts by Condition

Figure 3: In the treatment group, the difference between liberals and conservatives decreases after the chat conversation. A bootstrap analysis confirms that this difference is statistically significant.

if participants changed in polarization, they most often did so by moving one opinion step towards the center. Interestingly, the most radicalized participants (i.e. those who support Ukraine the most or the least) have similar rates of strong polarization change. Both the most and the least supportive participants are similarly likely to reduce their polarization by three opinion steps. At the same time, the most radicalized participants have different rates of small polarization changes. For the weakest depolarization (a move of one step towards the center), there is a difference between most and least supportive participants in the treatment group. Also, see figure A.4 in the appendix for a Sankey diagram of the opinion changes.

Figure 6 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

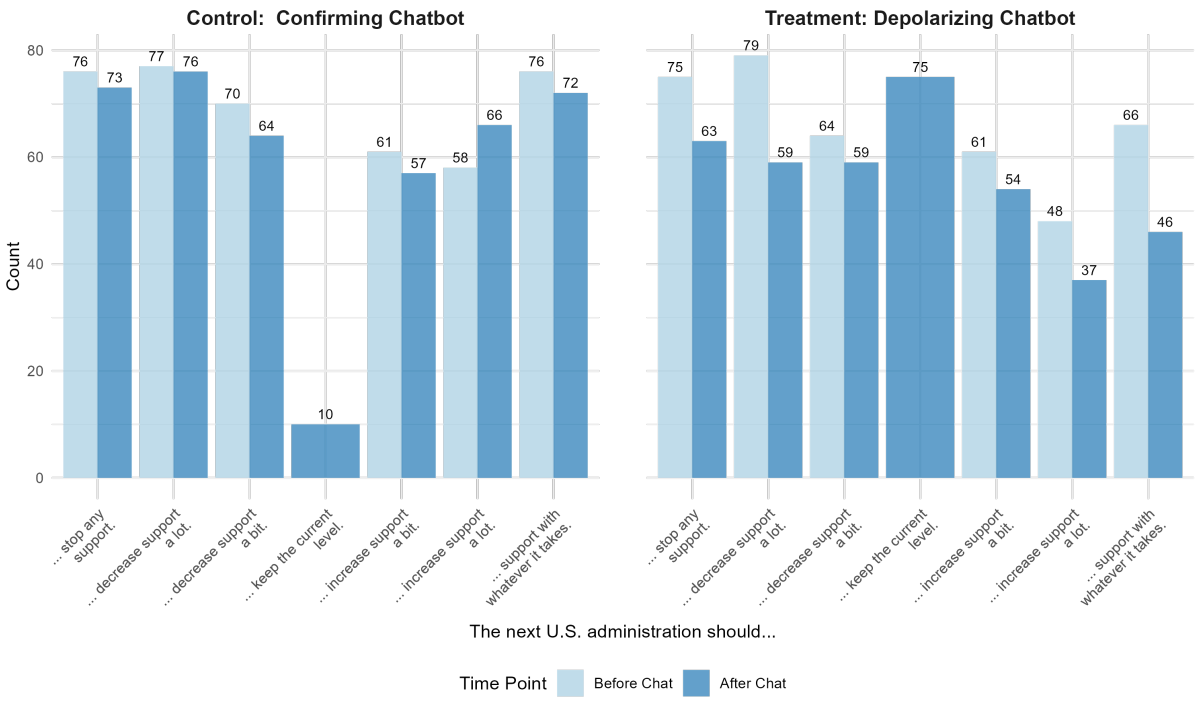


Figure 4: Bar plot of Opinion Counts by Condition. Note the difference between the treatment and control group in the center opinion "keep support at current level". Post chat, the number of participants who chose this option is 7.5 times higher in the treatment group than in the control group.

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 6 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 7 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?"

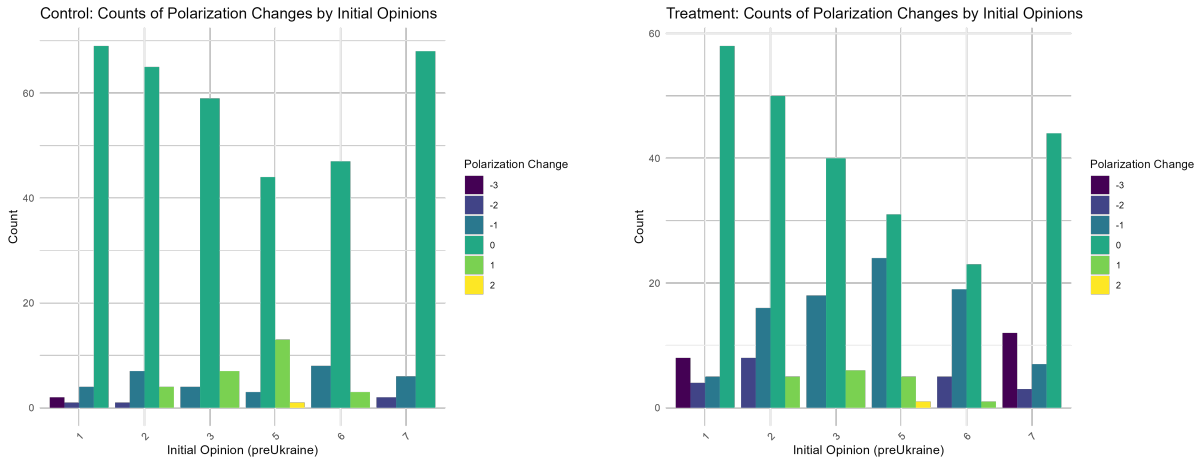


Figure 5: Political Polarization Changes by Initial Opinion for treatment and control. The most radicalized participants have similar rates of strong polarization change. Both the most and the least radical participants are similarly likely to reduce their polarization by three opinion steps.

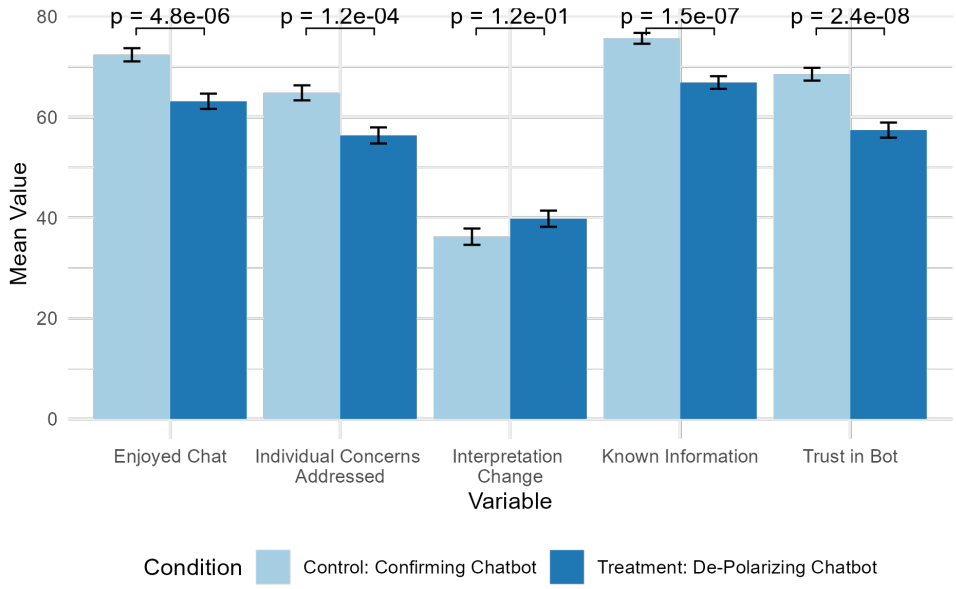


Figure 7: Mean values of enjoyment, trust and three measures of learning. Participants talking to the depolarization bot have enjoyed the chat less, felt that their individual concerns were addressed less and trusted the bot less. Participants felt that the depolarization bot provided them with more previously unkown information.

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 8 shows the results for the revealed preference outcomes. After the study endend, 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

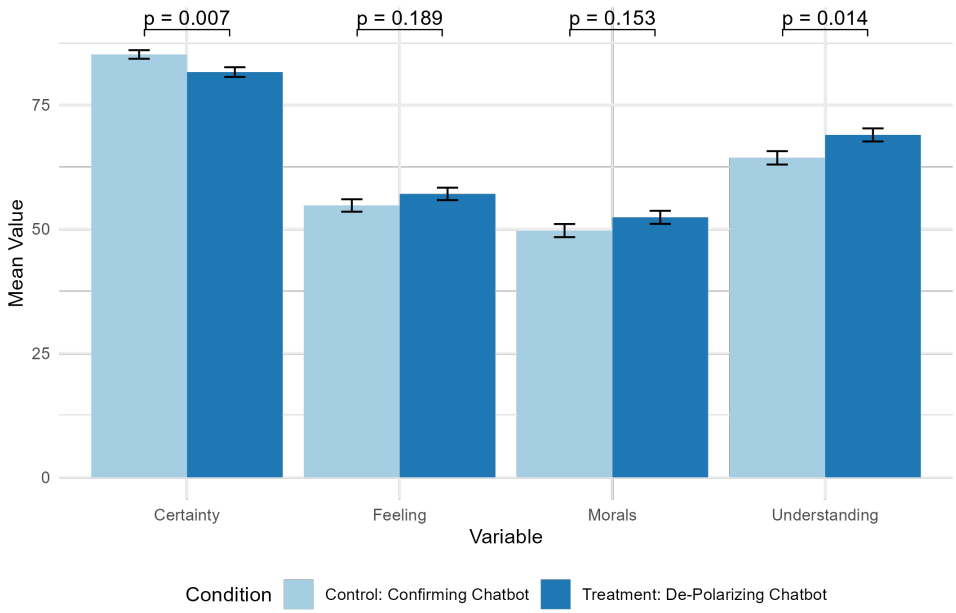


Figure 6: Mean values of cognitive uncertainty and affective polarization. The depolarization bot makes participants less certain. Affective polarization does not decrease in treatment, but the understanding for someone with a different opinion increases.

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.

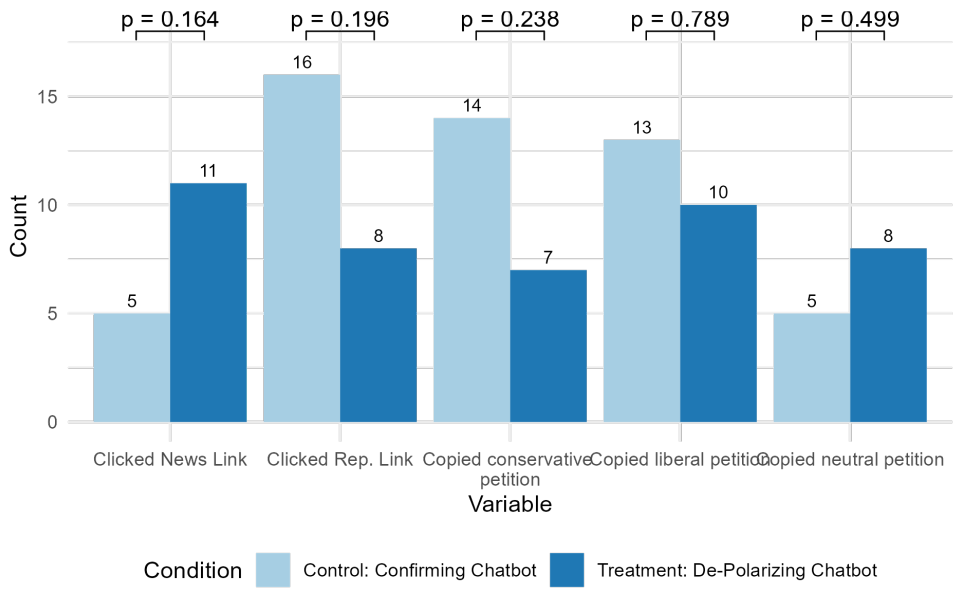


Figure 8: Click rates for the revealed preference outcomes. Due to very small click through rates, no difference is significant.

Figure 8 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

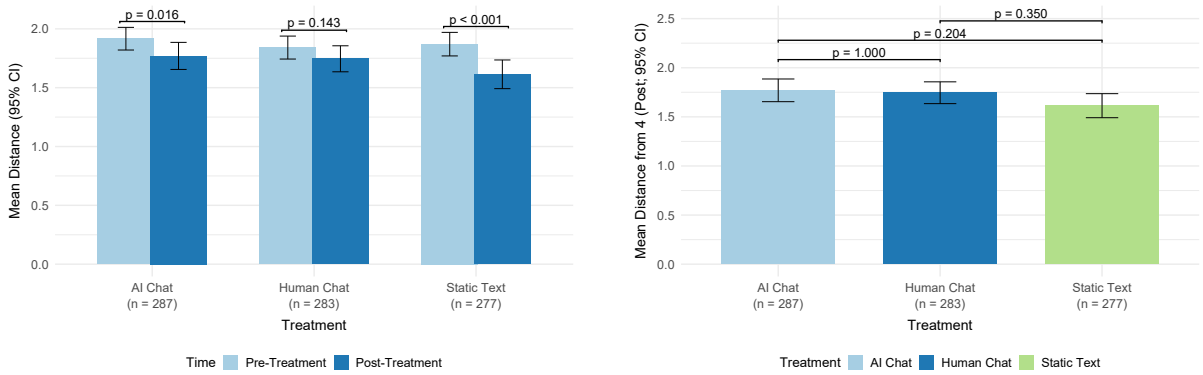
3.2.1 Effect on Ideological Polarization

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 9a 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 9: Average distances to the center answer option 4 ("In-between") on the question of immigration reduction. Panel (a) shows pre- and post-treatment distances for each treatment condition. Panel (b) compares post-treatment distances between treatment pairs. All three treatments show a significant decrease in distance to the center answer option 4 after the treatment, but none of the pairwise differences are significant; this indicates that all treatments worked but that none of them worked better than the others.

Table 8 in the appendix A.5.2 shows the pre-post changes seen in Figure 9a numerically.

Figure 9b shows the pairwise comparisons of the post-treatment distances between the treatments. No pairwise average treatment effect difference is significant.

Table 2: 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 2 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

$$\text{distance}_{it} = \alpha_i + \beta\,\text{Post}_t + \varepsilon_{it}, \tag{1}$$

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.

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

Table 3: 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

has a higher post-treatment distance than the second. p-values are from Welch two-sample t-tests with Bonferroni adjustment for multiple comparisons, matching the inference used in the figure. Statistical significance should be judged with these adjusted p-values; for example, none of the pairwise differences above is statistically significant at conventional levels.

Table 4: 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 4 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 in parentheses via the Std. Error column, with z statistics and p-values to assess significance.

Coefficients represent adjusted differences in the post outcome, holding controls fixed. Positive treatment coefficients indicate higher post-treatment distance than Static Text; negative coefficients indicate lower distance. In these results, both AI Chat

and Human Chat show small positive differences relative to Static Text (about 0.13) that are marginally significant at conventional levels. Baseline distance is strongly and positively associated with the post outcome, consistent with persistence in opinions. Age is negatively associated with distance, while English proficiency is not statistically different from zero. The female indicator is marginal and not conventionally significant at the 5% level. The simplified ethnicity, education, party, and region indicators are not statistically distinguishable from zero here. The "learned in chat" variable is positively associated with post distance; because it is measured after treatment, this association should be interpreted descriptively rather than causally. The model explains roughly 48% of the variation in the post outcome across 830 observations.

3.2.2 Effect on Affective Polarization and Opinion Conviction

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 9 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 10. 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.

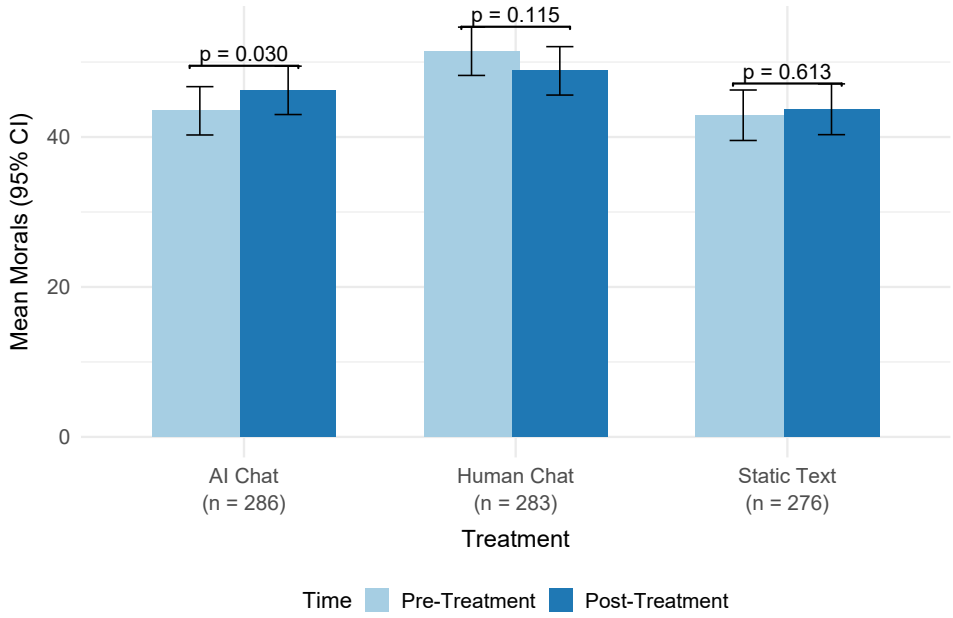


Figure 10: Treatment effects on the affective polarization measure: On a scale from 0 (Disagree completely) to 100 (agree completely), to what extent do you disagree or agree with this: "People with a very different opinion from mine on immigration, have the same moral values as me"?

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 11 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.

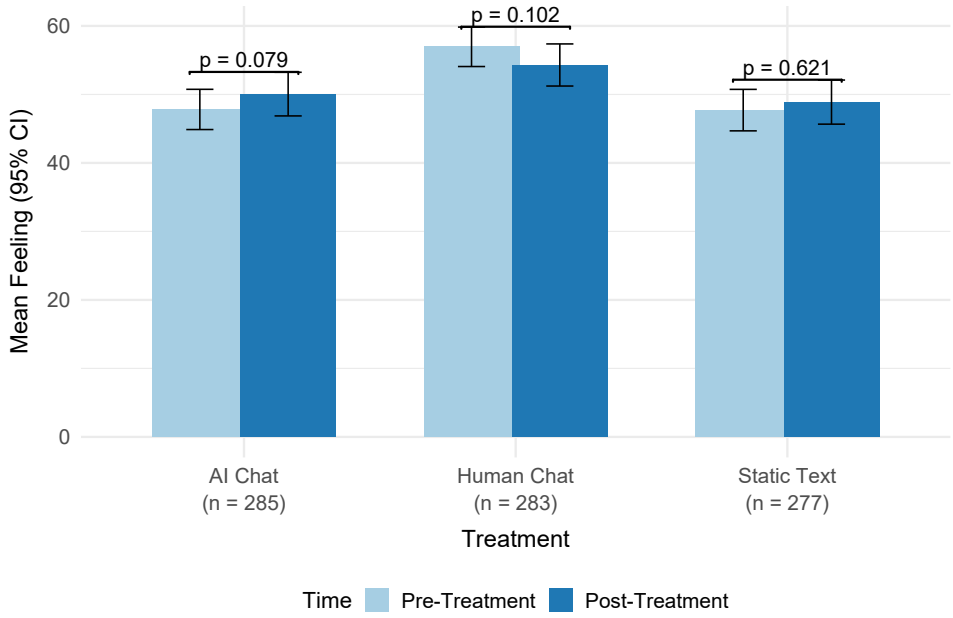


Figure 11: Treatment effects on the affective polarization measure: On a scale from 0 (Strong dislike) to 100 (Strong like), how do you feel about people with a very different opinion from yours on this question?

Figure 12 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.

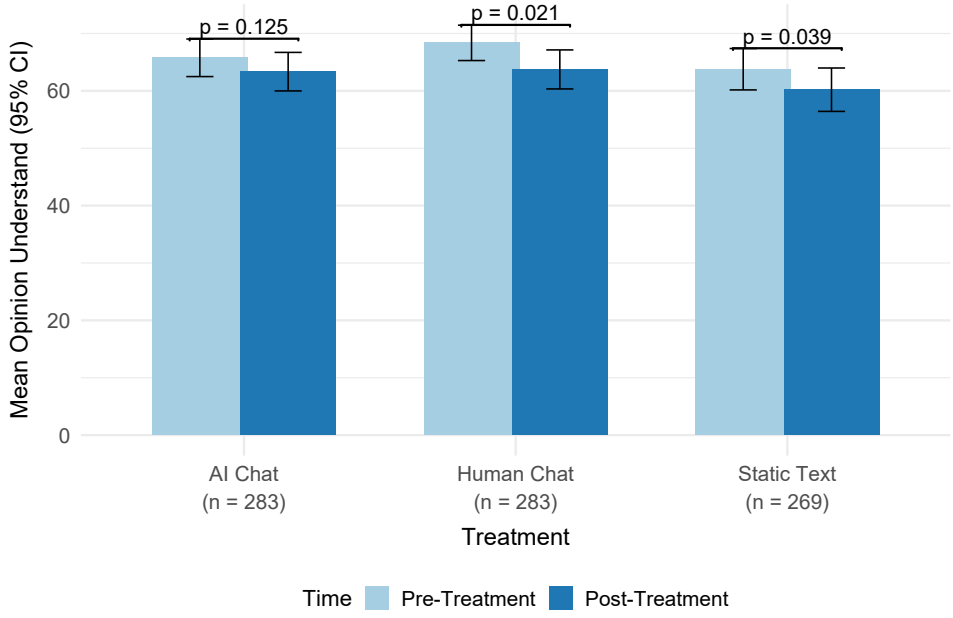
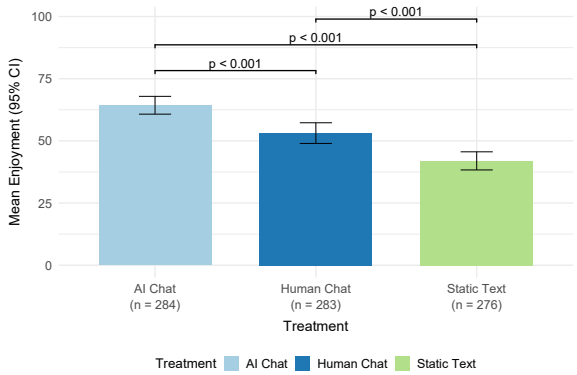
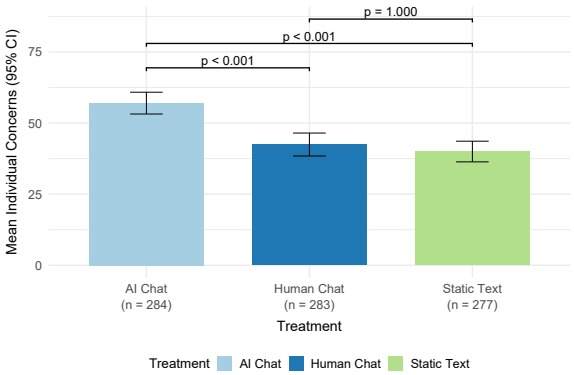


Figure 12: Treatment effects on the affective polarization measure: How well can you understand someone who has an opinion on this topic that is entirely different from yours?

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 13: Treatment effects on the enjoyment and individual concerns of the conversation.

This section compares the between treatment differences in two outcomes that are related to the on participant experience of the conversation: for the enjoyment outcome, 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 13a shows a comparision 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 13b 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

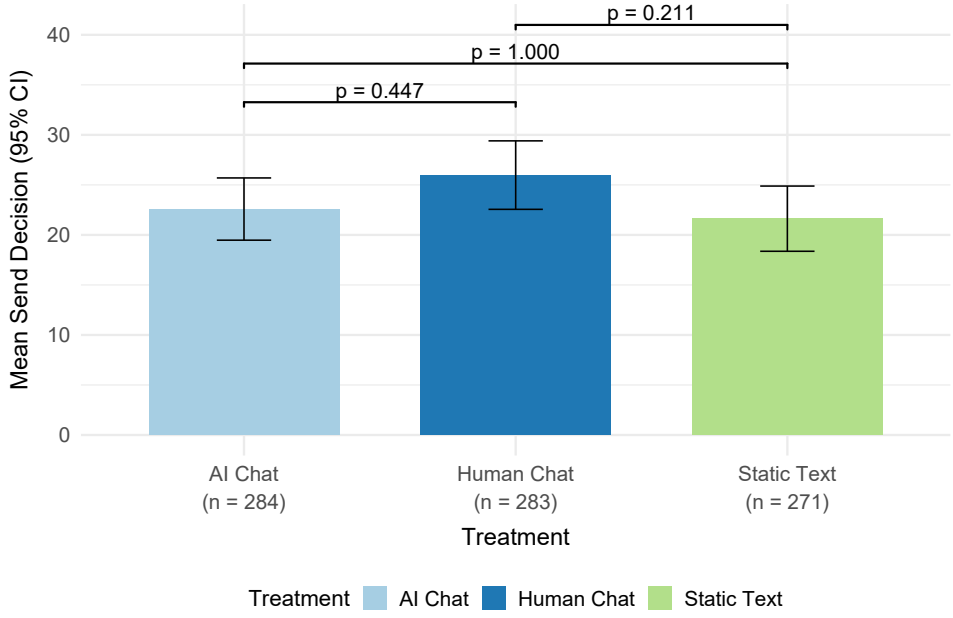


Figure 14: Treatment effects on the decision of how much money to send to a player with a opposite opinion in the dictator game.

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 14 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: Analysis of Arguments in the Chat

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 wasperceived 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 content analysis approach was implemented using GPT-4o. A predefined catalog of ten immigration-related arguments was developed, encompassing 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 struc-

tured prompt that instructed the model to function as an ”argument tagger,” evaluating whether each catalog argument appeared in the conversation and identifying any additional arguments not covered by the predefined catalog. 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 catalog with rationales for identification, (2) a list of catalog arguments present in the conversation, and (3) additional arguments expressed as 2-10 word phrases that captured distinct ideas not represented in the original catalog. This approach enabled comprehensive argument extraction while maintaining consistency across the 566 total conversations analyzed (292 human-to-human and 274 human-to-AI conversations).

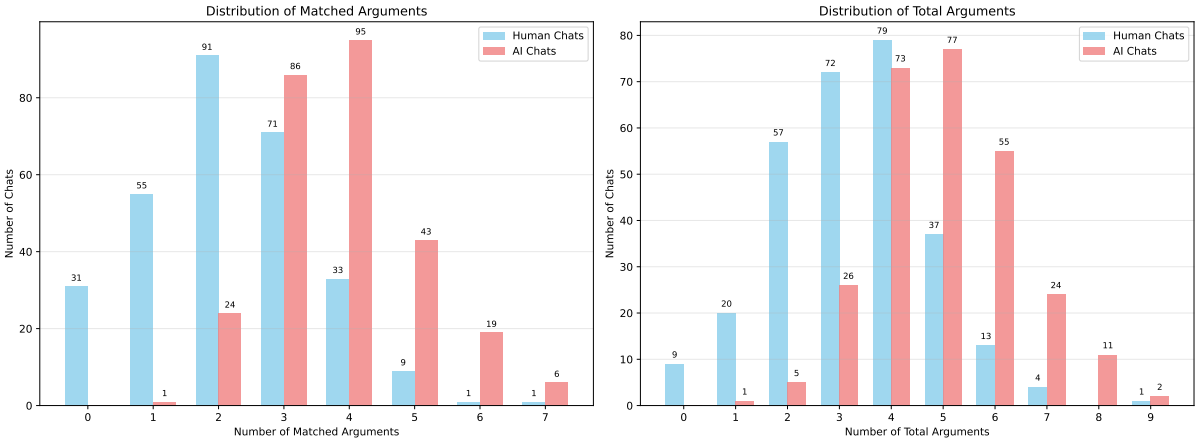


Figure 15: Distribution of Matched and Total Arguments by Chat Condition

Figure 15 shows two side-by-side bar charts comparing argument frequency between human-to-human and AI-to-human conversations. The left panel shows ”matched arguments” (arguments from a predefined catalog), while the right panel shows ”total arguments” (including both catalog and novel arguments). Key findings: AI conversations consistently produce more arguments than human conversations. For matched arguments, human chats peak at 2 arguments per conversation, while AI chats peak at 4 arguments. For total arguments, the difference is even more pronounced - human chats typically contain 3-4 total arguments, while AI chats frequently contain 5-6 arguments. Notably, some human conversations contain zero arguments, while this never occurs in AI conversations.

Table 5: Summary Statistics for Identified Arguments by Chat Type

Chat Type	Argument Type	Count	Min	Median	Max	Mean	Std	Total Arguments
Human	Matched	292	0	2.0	7	2.19	1.30	640
Human	Total	292	0	3.0	9	3.31	1.48	966
AI	Matched	274	1	4.0	7	3.86	1.15	1058
AI	Total	274	1	5.0	9	5.00	1.38	1370

To obtain these data, the chat conversations were processed using GPT-4o for automated argument identification. In the human-to-human condition, 292 chat conversations between human participants were analyzed, while in the human-to-AI condi-

tion, 274 conversations between human participants and an AI chatbot were examined. Each conversation was processed through GPT-4o using a structured prompt that identified arguments from a predefined catalog of 10 immigration-related arguments ("matched arguments") as well as additional arguments not covered by the catalog ("other arguments"). The "total arguments" represents the sum of matched and other arguments per conversation. The results show that AI-mediated conversations contained significantly more arguments per chat (median = 5.0 total arguments) compared to human-only conversations (median = 3.0 total arguments), with AI conversations also showing higher argument density across both matched catalog arguments and novel arguments identified by the language model. Statistical tests confirmed these differences are highly significant: Mann-Whitney U tests revealed significant differences for both matched arguments ($U = 13,846$, $p < 0.001$) and total arguments ($U = 16,291$, $p < 0.001$), indicating that AI-mediated conversations consistently generated more argumentative content than human-only discussions.

Table 6: 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 16 shows a comparison of the count of pro-immigration and con-immigration arguments by chat condition.

4 Discussion

Finding compromises lies at the heart of democratic processes and is a necessity to get any meaningful policy done. Unhealthy levels of polarization make compromises difficult. Having tools and processes that can reduce polarization can therefore be thought of 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 and that they are as effective as incentivized human persuaders or traditional information

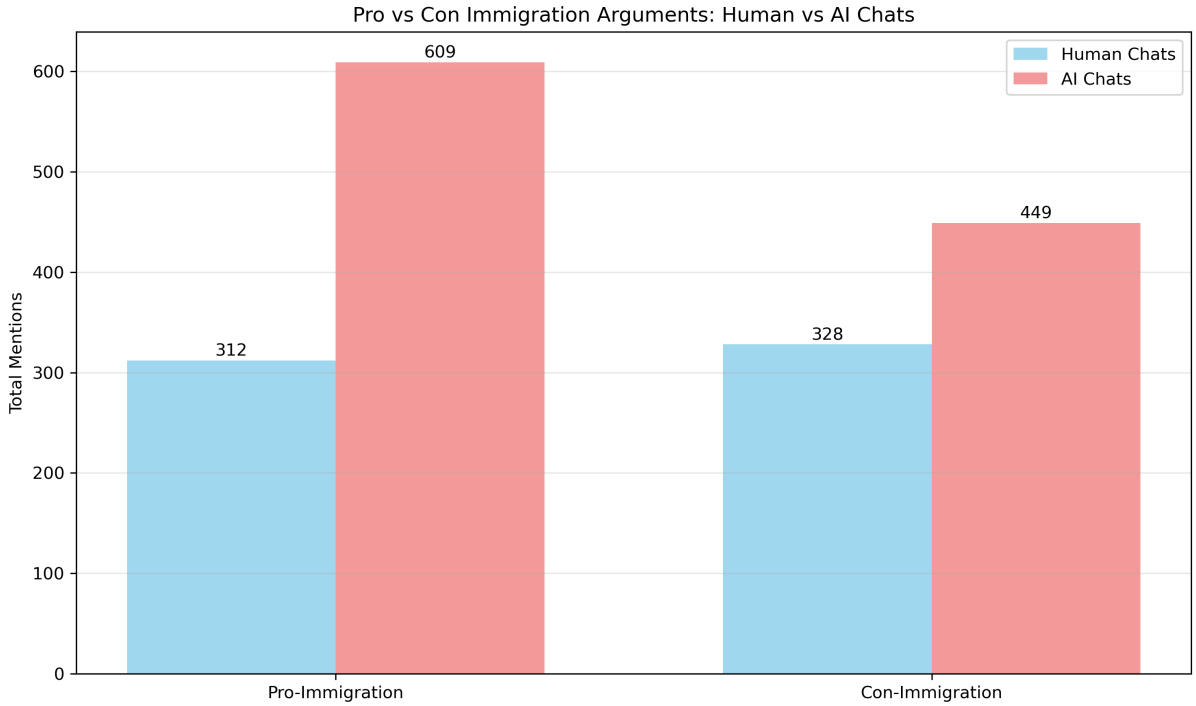


Figure 16: Count of Pro-Immigration and Con-Immigration Arguments by Chat Condition

interventions in reducing stated ideological polarization and have unique benefits, namely a improved enjoyment and an superior 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.

While the EU Digital Services Act (DSA) prohibits illegal content and misinformation, it does not adequately address the nuances of AI-powered persuasion that is grounded in factual information. This gap in regulation presents a challenge: if AI-driven bots can be utilized as scalable tools to persuade individuals without resorting to misinformation or illegal content, should they be permitted to operate freely?

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.

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 the first comprehensive experimental evidence that AI-powered conversational agents can effectively reduce political polarization across multiple contentious issues. Through two pre-registered randomized controlled trials with representative samples of the U.S. population, I demonstrate that carefully designed AI interventions can successfully persuade participants to adopt more moderate positions, with important implications for both the application and regulation of AI in political discourse.

The first experiment (N=811) established that an AI chatbot could persuade participants to moderate their views on U.S. support for Ukraine, reducing overall ideological polarization by approximately 20 percentage points. A positive effect persisted in a follow-up study conducted one month later. The second experiment (N=838) provided comparative insights by testing AI persuasion against incentivized human persuaders and static text on immigration policy. All three interventions significantly reduced participants' distance from moderate positions when compared to their own pre-treatment distributions. However, the between-treatment comparisons revealed no statistically significant differences in persuasive effectiveness.

Across both experiments, the effect of AI persuasion with respect to reducing measures of affective polarization was more modest and in the second experiment there was no effect on a revealed preference outcome.

Overall, this research provides evidence that AI-powered persuasion can be used to effectively influence political polarization. As a comparatively cheap and scalable tool, this suggests both positive uses cases to reduce polarization but also serves as a cautionary tale about the potential for AI manipulation.

References

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

Argyle, Lisa P et al. (2025). "Testing theories of political persuasion using AI". In: Proceedings of the National Academy of Sciences 122.18, e2412815122.
Arieli, Itai and Yakov Babichenko (2019). "Private bayesian persuasion". In: Journal of Economic Theory 182, pp. 185–217.
Bai, Hui et al. (2025). "LLM-generated messages can persuade humans on policy issues". In: Nature Communications 16.1, p. 6037.
Belot, Michèle and Guglielmo Briscese (2022). Bridging America's Divide on Abortion, Guns and Immigration: An Experimental Study. Tech. rep. CEPR Discussion Papers.
Boxell, Levi, Matthew Gentzkow, and Jesse M. Shapiro (2022). "Cross-Country Trends in Affective Polarization". In: The Review of Economics and Statistics 104.5, pp. 981–1001. DOI: 10.1162/rest_a_01160. URL: https://doi.org/10.1162/rest_a_01160.
Brown University (2020). U.S. is polarizing faster than other democracies, study finds. Accessed: 2024-12-06. URL: https://www.brown.edu/news/2020-01-21/polarization.
Brown, Jacob R et al. (2023). "The increase in partisan segregation in the United States". In: Nottingham Interdisciplinary Centre for Economic and Political Research Discussion paper 2023-09.
Callander, Steven and Juan Carlos Carbajal (2022). "Cause and effect in political polarization: A dynamic analysis". In: Journal of Political Economy 130.4, pp. 825–880.
Castiglioni, Matteo et al. (2020). "Online bayesian persuasion". In: Advances in neural information processing systems 33, pp. 16188–16198.
Chen, Daniel L, Martin Schonger, and Chris Wickens (2016). "oTree—An open-source platform for laboratory, online, and field experiments". In: Journal of Behavioral and Experimental Finance 9, pp. 88–97.
Costello, Thomas H, Gordon Pennycook, and David G Rand (2024). "Durably reducing conspiracy beliefs through dialogues with AI". In: Science 385.6714, eadq1814.
Fafchamps, Marcel et al. (2024). Diffusion in social networks: Experimental evidence on information sharing vs persuasion. Tech. rep. National Bureau of Economic Research.
Jacobs, Julian (2024). "The artificial intelligence shock and socio-political polarization". In: Technological Forecasting and Social Change 199, p. 123006.
Kamenica, Emir (2019). "Bayesian persuasion and information design". In: Annual Review of Economics 11.1, pp. 249–272.
Kamenica, Emir and Matthew Gentzkow (2011). "Bayesian persuasion". In: American Economic Review 101.6, pp. 2590–2615.
Kempfxd, Elisabeth and Margarita Tsoutsoura (2024). "Political polarization and finance". In: Annual Review of Financial Economics 16.
Kerr, John, Costas Panagopoulos, and Sander Van Der Linden (2021). "Political polarization on COVID-19 pandemic response in the United States". In: Personality and individual differences 179, p. 110892.
Mill, Wladislaw and John Morgan (2022). "The cost of a divided America: an experimental study into destructive behavior". In: Experimental Economics 25.3, pp. 974–1001.
Schoenegger, Philipp et al. (2025). "Large Language Models Are More Persuasive Than Incentivized Human Persuaders". In: arXiv preprint arXiv:2505.09662.
Schwartzstein, Joshua and Adi Sunderam (2021). "Using models to persuade". In: American Economic Review 111.1, pp. 276–323.

Sunstein, Cass (2018). # Republic: Divided democracy in the age of social media. Princeton university press.
Voelkel, Jan G et al. (2024). "Megastudy testing 25 treatments to reduce antidemocratic attitudes and partisan animosity". In: Science 386.6719, eadh4764.
Wang, Yun (2015). Bayesian persuasion with multiple receivers. SSRN.

A Appendix: Additional Materials

A.1 Distribution of Polarization Changes by Condition

Table 7: 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	2	0.5%	2	0.5%
Sum	418	100%	393	100%
Participants in control and treatment	51.5%		48.5%	

A.2 Main Survey Question

The main question was: How much do you agree with the following statement: "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.3 AI System Prompts

A.3.1 System Prompt for Treatment Group

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

> ”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.”
> 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, es-

pecially 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.3.2 System Prompt for Control Group

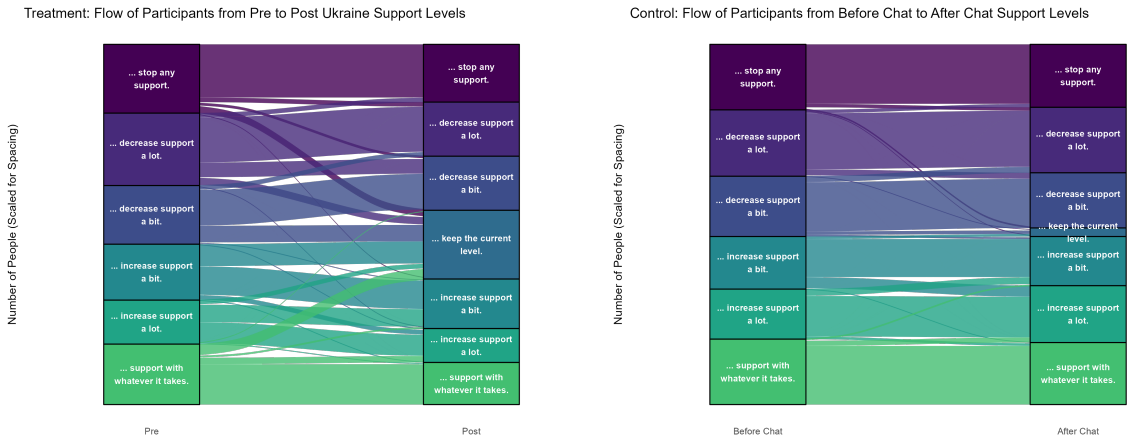
The system prompt for instructing the AI in the control group was:

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

A.4 Sankey Graphs



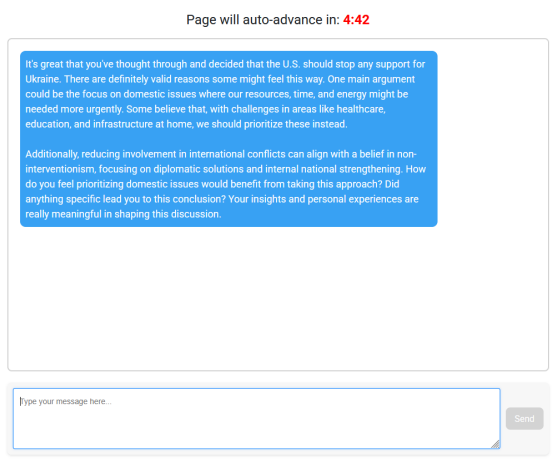
A.5 Chat Interface

A.5.1 Can AI predict the persuasion success based only on the conversation?

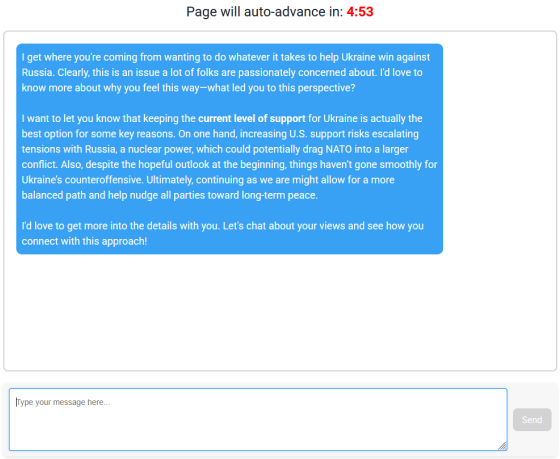
The previous section documented that the gpt-4o model can successfully persuade some participants to change their opinion. Is it possible for the same model to also directly predict whether a participant changed their opinion based on the conversation? Providing a prediction of opinion change could be 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.



(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 17: Example chat conversations from the experiment. 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.

Table 8: 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. $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

A.5.2 Experiment 2: Pre-Post Change in Distance from Center (4) by Treatment

Table 8 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.

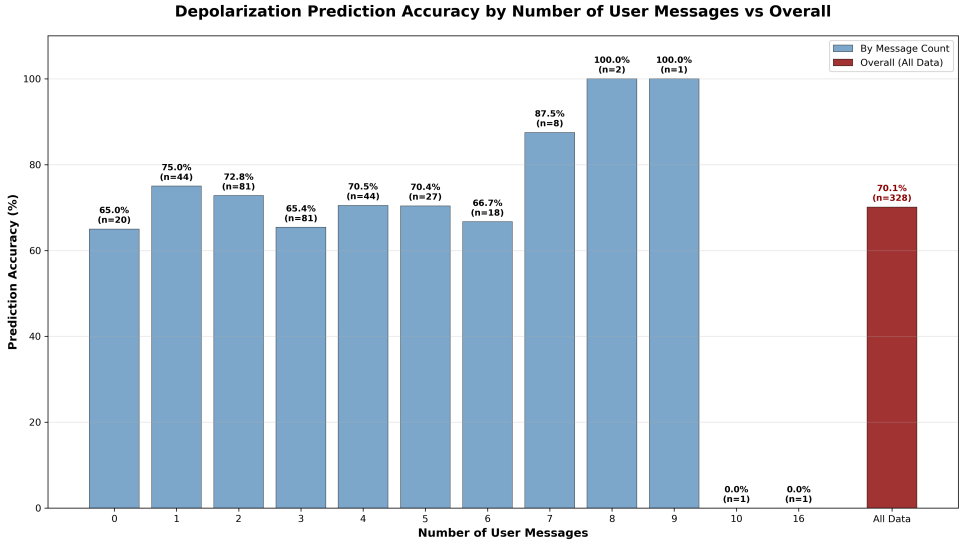


Figure 18: Average prediction accuracy of the gpt-4o model.

A.5.3 Experiment 2: Numerical Summary of Treatment Effects on Affective Polarization and Opinion Conviction

Table 9 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.

A.5.4 Experiment 2: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
- **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
- **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

Table 9: 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)

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

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.

A.5.5 Experiment 2: 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.

A.5.6 Experiment 2: Chat Analysis

Table 10: 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**

A.5.7 Distribution of Opinions on Immigration Reduction before Screening Out Initially Depolarized Participants

Figure 19 shows the distribution of pre-opinions on Ukraine support.

Figure 19 shows the distribution of pre-opinions on immigration reduction.

A.5.8 Experiment 2: Random Sample of "Other" Arguments

- Immigrants contribute significantly to tax revenues, including income, payroll, sales, and property taxes.

Table 11: 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**

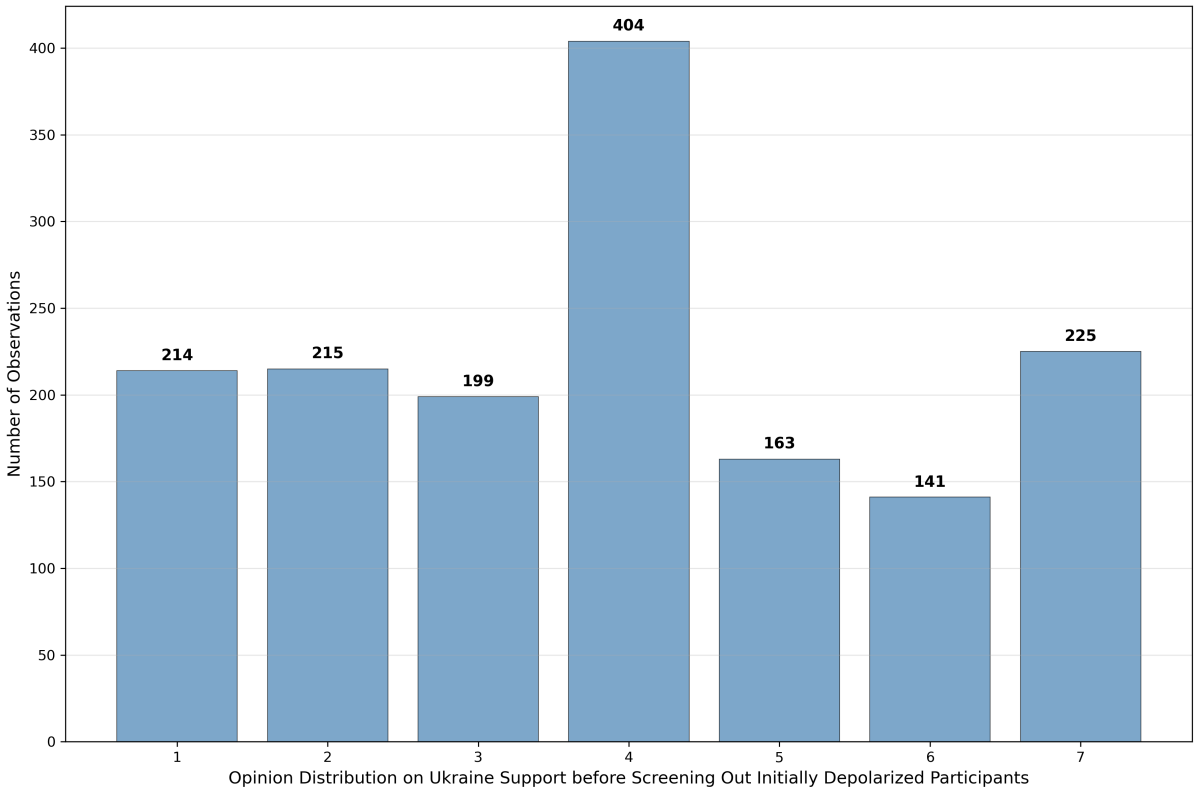


Figure 19: Distribution of Pre-Opinions on Immigration Reduction

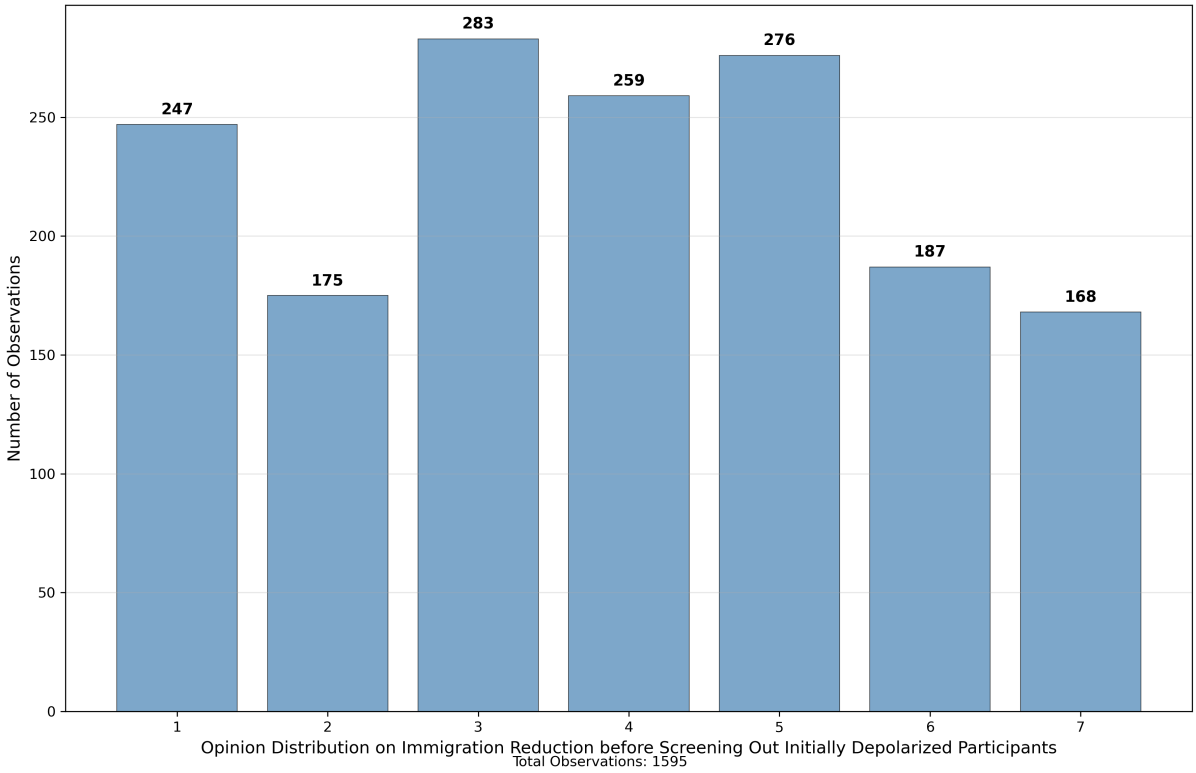


Figure 20: Distribution of Pre-Opinions on Immigration Reduction

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

Table 12: 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