Title: AI Chat Assistants can Improve Conversations about Divisive Topics

Authors: Lisa P. Argyle,[1] Ethan C. Busby,[1] Joshua Gubler,[1] Chris Bail,[2*] Thomas Howe,[3] Christopher Rytting,[3] David Wingate[3]

[1]Department of Political Science, Brigham Young University,
[2]Department of Sociology, Political Science, and Public Policy, Duke University,
[3]Department of Computer Science, Brigham Young University

[*]To whom correspondence should be addressed; E-mail: christopher.bail@duke.edu.

One Sentence Summary: In a large field experiment, we show an AI chat assistant can improve quality and reduce divisiveness in conversations about gun control.

Abstract: A rapidly increasing volume of conversation occurs online, but divisiveness and conflict can fester in digital interactions. Such toxicity increases polarization and corrodes the capacity of diverse societies to cooperate in solving social problems. Scholars and civil society groups promote interventions that make conversations less divisive or more productive, but scaling these efforts to online discourse is challenging. We conduct a large-scale experiment that demonstrates how online conversations about divisive topics can be improved with artificial intelligence tools. Specifically, we employ a large language model to make real-time, evidence-based recommendations intended to improve participants' perception of feeling understood. These interventions improve reported conversation quality, reduce political divisiveness, and im-

prove the tone, without systematically changing the content of the conversation or moving people's policy attitudes.

Main text:

Background and Theory

Most of the world's population now employs the internet to converse with others. More than 100 billion messages are sent every day on Facebook and Instagram alone (*1*), and approximately 7 billion conversations occur daily on Facebook Messenger (*2*). Such conversations can have far-reaching impact. Some of the largest social movements in human history have recently emerged out of sprawling conversations on social media, and discussions between high profile social media users can shape the stock market, politics, and many other aspects of human experience (*3–6*).

Social scientists have long observed that conversation is the "soul of democracy" insofar as it helps diverse groups of people identify solutions to shared problems (*7–11*). However, there is growing concern about the quality of discourse in online settings (*12–15*). Nearly half of social media users report observing mean or cruel behavior, and many indicate that divisiveness and incivility complicate a variety of relationships in their lives – with family, friends, and work colleagues (*16*). As such, many members of the public either avoid online discussions about politics or unwittingly find themselves arguing online in a corrosive, unconstructive manner (*17–20*). Divisive rhetoric has been linked to partisan violence (*21, 22*), disengagement from politics and public life (*20, 23*), and reduced capacity to find compromise (*24*).

Meanwhile, scholarship on how to facilitate more productive conversations has grown alongside such concerns (*25–33*). These studies work to identify a range of strategies such as active listening, validation of opposing views, and value-based messaging that increase the likelihood

that members of rival groups find common ground. These tactics help address the divisions and polarization that can otherwise occur in social interactions between people (*34*). Though such strategies often do not result in immediate political compromise, many scholars believe that improving the quality of political discourse will have broader benefits related to social cohesion and democracy (*19, 27, 35*). That is, "hearing the other side" (*25*) can be productive even if disagreement remains. Such dialogue is a necessary, even if insufficient, condition for increasing mutual understanding, compromise, and coalition building.

In what follows, we present the results of a field experiment that employs cutting-edge artificial intelligence tools – in this case, the large language model (LM) GPT-3 – to *scale up* evidence-based interventions to reduce the divisiveness of politically charged conversations. We invited proponents and opponents of gun regulation in the United States into online conversations, randomly assigning a chat assistant powered by GPT-3 to some of the participants. We show that intervention by the chat assistant, which recommends real-time, context-aware, and evidence-based ways to rephrase messages, improved perceptions of conversation quality, decreased divisiveness, and reduced toxicity in the chats, primarily for the partner of the person who used the AI assistant.

AI Tools in Social Science

Political actors and social scientists increasingly use artificial intelligence tools to influence and study the social world (*36–39*). Language models, such as the prominent and recently released ChatGPT, highlight the ability of artificial intelligence to generate human-sounding text and perform tasks previously thought impossible (*40*). Given their potential to identify and replicate complex patterns in text, LMs provide a promising new way to explore social outcomes (*41*). One important advance of these models is their capacity for "few-shot" learning, wherein they learn to perform a task from just a few exemplars without requiring parameter updates (*42*).

While many observers are rightfully concerned about the negative effects of biases present in LMs and other AI tools (*43–47*), the same model features that generate these biases also enable LMs to produce text that is nuanced and multifaceted in its representation of a range of people, tones, ideas, and attitudes (*41*). Prior AI-in-the-loop applications have demonstrated that AI can help people be more empathetic in peer mental health support conversations (*48*), and that inducing reflection and restatement can improve the quality of divisive conversations (*49, 50*). We build on that work to demonstrate that dynamic, real-time, and context-aware AI-generated recommendations can improve the quality of political conversations and reduce political divisiveness.

Feeling Understood in Conversation

In this research project, we specifically define "better quality" conversations as those in which people have an increased perception that they are *better understood by the person with whom they are talking*. Although all conversations across lines of difference do not reduce divisiveness (*34, 51*), the feeling of being understood has been shown to generate a host of positive social outcomes (*30,33,52–54*). Importantly, research shows that the benefits of such conversations *do not require persuasion* or agreement between participants on the issues discussed, just a feeling that each person's perspective was heard, understood, and respected. As such, our use of AI in this experiment does not seek to change participants' minds; we suggest this as a model for how AI can be employed without pushing a particular political or social agenda. Our focus on feeling understood is also a response to some concerns about over-emphasizing increasing civility and specific forms of depolarization as normative goals (*10, 23, 55*).

Research suggests a number of specific, actionable conversation techniques to effectively increase the perception of being understood (*30–32*), which are used worldwide (*35*). In practice, civil society or academic approaches typically have trained moderators and instructors teach,

model, and provide practice in developing respectful conversation skills. While effective, these interventions reach only a small fraction of those caught in divisive conversations daily. The challenge is implementation at scale: helping individuals recognize and remember how to apply these techniques in real-world conversations, and/or find the will to apply them in the moment of a (heated) conversation.

Hypotheses

We developed an AI chat assistant to fill this need and act as a real-time moderator. The assistant makes repeated, tailored suggestions on how to rephrase specific texts in the course of a live, online conversation, without fundamentally affecting the content of the message. The suggestions are based on three specific conversation-improving techniques from the literatures mentioned earlier: **restatement**, simply repeating back a person's main point to demonstrate understanding; **validation**, positively affirming the statement made by the other person without requiring explicit statements of agreement (e.g. "I can see you care a lot about this issue"); and **politeness**, modifying the statement to use more civil or less argumentative language.

Our pre-registered expectations are that individuals in chats with political opponents where one participant has the rephrasing assistance of our AI tool will report higher conversation quality, hold less derisive opinions of their political opponents and express a greater willingness to believe that their opponents have valid ideas, even if they disagree, than those in untreated conversations. We expect no treatment effect on change in policy attitudes. We also expect treatment exposure will reduce toxicity and improve the tone of downstream non-treated messages.

Study Design

We test these hypotheses in an online chat experiment about gun regulation. Gun policy is the subject of the conversations because it has near constant salience in American social and political life and is a sharply divisive issue (*56–59*). Respondents first completed a short survey, which ended with a summary measure of their feelings about gun policy in the United States. They were then routed to our custom-built online chat platform where they were matched with another study participant who disagreed with them on gun regulation. Once matched, conversation pairs were randomly assigned to the treatment or control condition, and partners proceeded to have a conversation. In the treatment condition, one partner received a rephrasing prompt for the first message longer than four words in every other conversation turn, regardless of the specific tone or content of the message. Figure 1 shows how the rephrasing prompts from GPT-3 fit into the conversational flow. Participants could choose to send one of three AI-suggested alternatives, their original message, or edit any message. Prior to the start of the conversations, individuals assigned to receive suggestions from the LM were shown a brief tutorial to orient them to the rephrasing process.

After completing the chat, respondents were routed to another survey that measured their impressions of their conversations, levels of divisive attitudes towards those who disagree with them on gun regulation, and the same measures of their views of gun regulation as in the pre-chat survey. Nearly 3 months after the experiment, we recontacted participants to explore the durability of our treatment effects; in line with the broader literature on these kinds of interventions (*60*), we found no evidence for the persistence of our treatments.

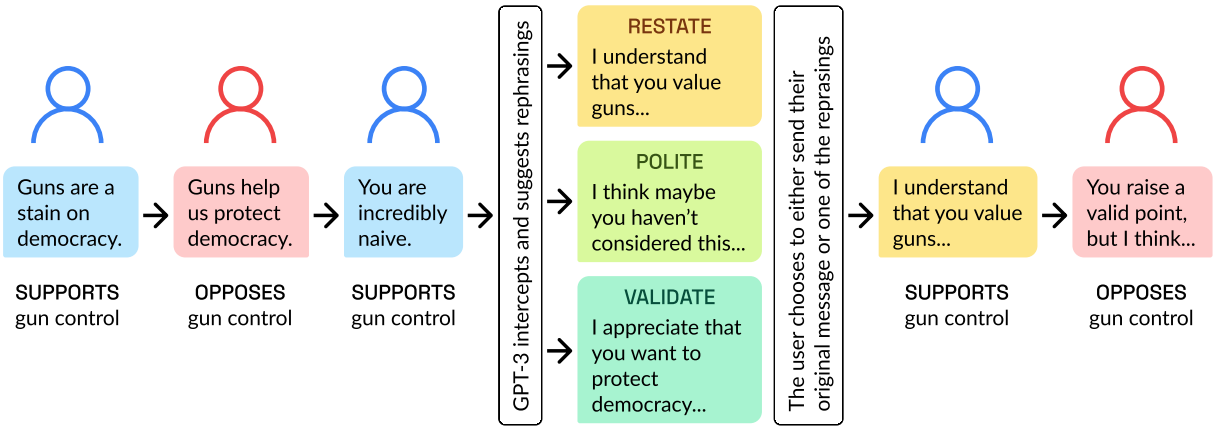

Figure 1: **Treated Conversation Flow**: Respondents write messages unimpeded until, at pre-defined intervals, the chat assistant intercepts the treated user's message, using GPT-3 to propose evidence-based alternative phrasings, while retaining the semantic content. It suggests three randomly ordered alternatives to the author of the message, and presents the opportunity to accept or edit any of these rephrasing suggestions or send their original message. Their choice is sent to their partner and the conversation continues.

Results

In October 2022, 1,574 people completed participation in our field experiment. By design, conversations were expected to continue until the treated individual in the chat received four rephrasing prompts; equivalently, control conversations were set to finish after one partner would have received four interventions, had they been provided. However, in practice, only 698 (44%) of participants were in chats that lasted the full intended length (see the Supplemental Materials for further discussion). On average, 12 total messages were sent in each conversation with a total of 2,742 AI rephrasings suggested. The AI-suggested rephrasings were accepted by chat participants two-thirds (1,798) of the time. Accepted messages were roughly evenly split between restate (30%), validate (30%), and polite (40%) styles.

Rephrasing Quality

We first verify that the chat assistant functioned as intended. We find that compared to the original message sent, the AI alternatives selected by participants had improved politeness, tone, and other textual qualities (see results in the Supplemental Materials). To confirm that these suggestions changed the tone without altering the substantive content of messages, we analyzed all 10,695 messages sent with more than 4 words. Figure 2 presents the results of an an automated pipeline using a variety of ML techniques to explore this point. In it, messages are embedded in a 2D space and clustered; we used GPT-3 to automatically generate a short summary of the content of each cluster, shown in the figure legend. Specific details on the generation of this figure can be found in Supplementary Materials.

Panel (A) of Figure 2 shows the topic clusters and corresponding GPT-3-generated labels. The labels show that the vast majority of messages sent on the platform were on-topic; additional manual checking confirmed this. Panel (B) colors the same points as either "normal" or "rewritten". As can be seen, the AI-rewritten messages are spread evenly throughout the semantic space. This indicates that rewritten messages did not fundamentally alter topical distribution, nor were there obvious degeneracies (such as mapping all rewritten points to a single cluster, or creating fundamentally new clusters).

Treatment Effects

Like many field experiments, ours faces a treatment dosage challenge: many participants assigned to be "treated" (to receive four interventions) only received partial treatment (fewer than four interventions, including zero). To avoid biased estimates as a result of simple post-treatment conditioning on number of received interventions (*61*), we use three different approaches to measure treatment effects, estimating intent-to-treat (ITT) effects and two measures of the complier average causal effects (CACE):

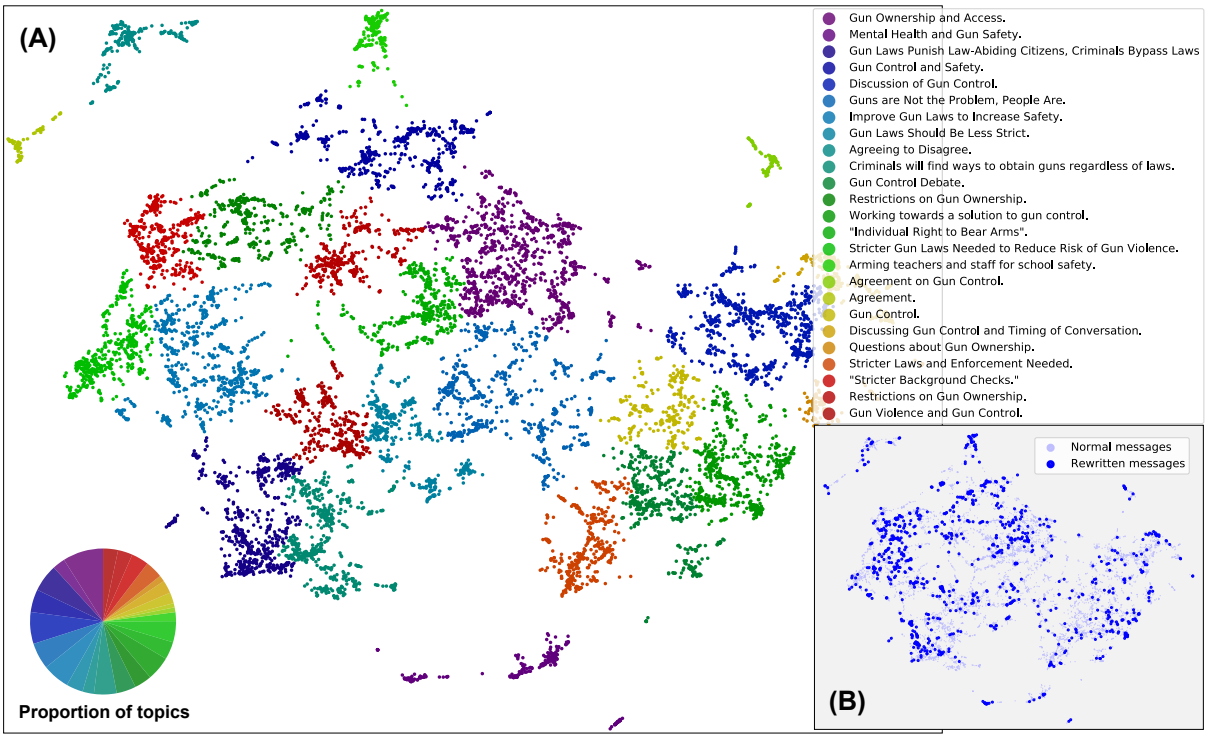

Figure 2: **Message Content Clusters**, as named by GPT-3. **Panel A** presents a visualization of the topical distribution of messages sent on the platform. Each point is a message; messages are clustered and colored by semantic similarity and automatically labeled by GPT-3 (legend titles). **Panel B** shows the distribution of the rephrased messages (larger, darker points) across the topic space.

(1) ITT: the most conservative measure of our treatment effects, ITT calculates effects based on random *treatment assignment* alone and ignores different treatment exposure levels.

(2) Placebo-controlled CACE: because participants in both the treatment and control conditions have a conversation in which early departure is equally possible, we calculate the treatment effect in subgroups of the study population based on the number of interventions they received (or would have received, had they not been in a placebo conversation).This method is causally identified for each subgroup of the data under the assumption that the treatment – rephrasing interventions – is unrelated to people's persistence in the conversation (see the Supplementary Materials for several tests supporting this assumption).

(3) Two-Stage Least Squares CACE: these models use *treatment assignment* as an instrument for the extent of *treatment exposure*. Additionally, because treatment exposure depends on a conversation actually occurring, we include as an additional instrument an indicator for whether the subject's partner sent a single message (a decision made after assignment but before any treatment exposure). We also include controls for all variables that are significant ($p < 0.1$) predictors of conversation length: pre-treatment gun control position, party ID, race, education, and employment status. We model three definitions of compliance; treating any exposure as full compliance can be interpreted as a lower bound of the true effect, whereas treating only full exposure as compliance might be considered an upper bound (*62*). For reasons described in the Supplementary Materials, we believe the placebo-control CACE estimates are a better estimate of the true effect, but provide these as a more conservative effect estimate.

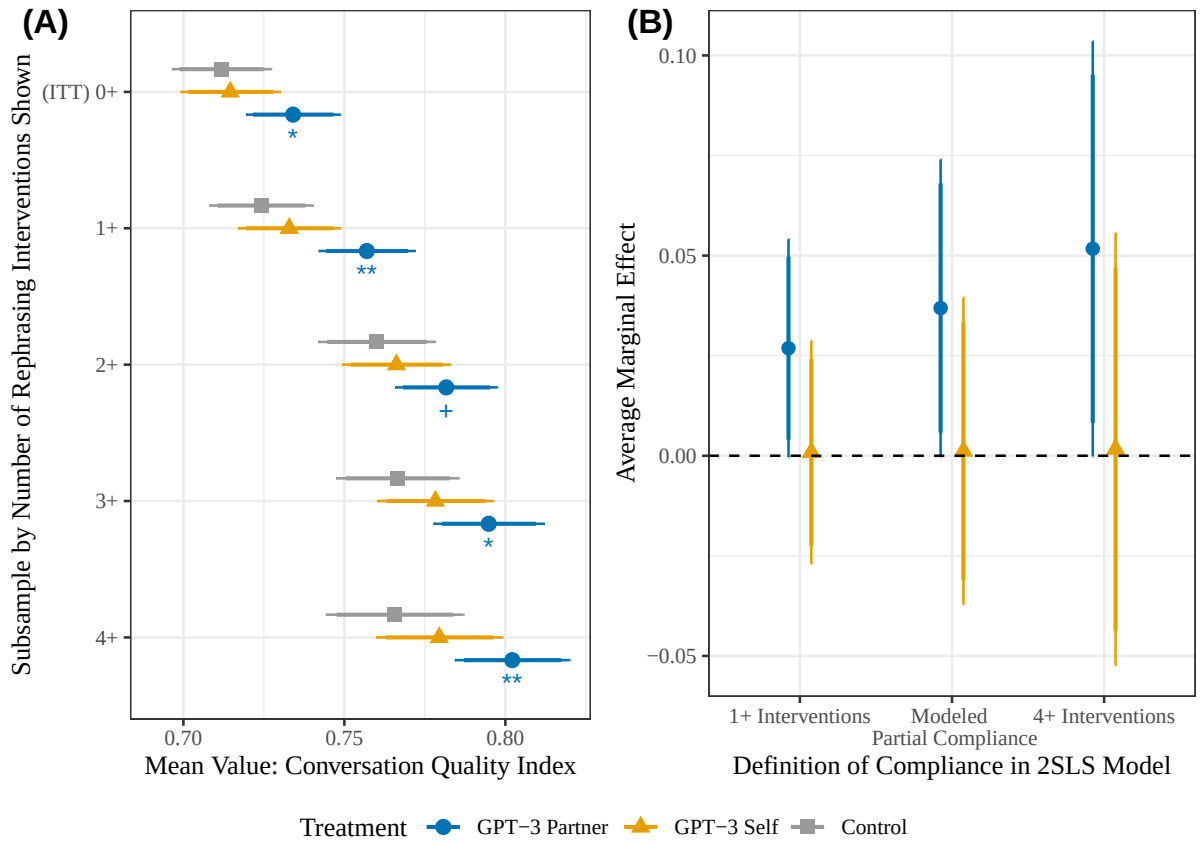

Figure 3: **Analysis of conversation quality index**. The index is scaled from 0 (lowest level of quality) to 1 (highest quality on all measures). **Panel A**: ITT and placebo-controlled CACE estimates. The number of rephrasing interventions are overlapping sets, such that 0+ includes all observations. Presents means, 90%, and 95% confidence intervals based on unadjusted standard errors. **Panel B**: Two-Stage Least Squares CACE estimates. Treatment assignment and whether the partner sent a single message are instruments for compliance, plus demographic controls. Y-axis shows average marginal effect of compliance, with standard errors clustered at the conversation level. The left model defines compliance as seeing 1 or more rephrasing interventions (lower bound estimate), the right model defines compliance as seeing 4 or more rephrasing interventions (upper bound estimate), and the middle model defines partial compliance (1 intervention = .25, 2 = .5, 3 = .75, 4+ = 1). Treatment increases the conversation quality reported by the partner of the person receiving interventions.$+p < .1, *p < .05, **p < .01, ***p < .001$

Random assignment occurs at the conversation level, but only one person in the conversation receives the chat assistant intervention. Therefore, we estimate three effects: one for the person who used the assistant ("GPT-3 Self"), one for those whose partner used the assistant ("GPT-3 Partner"), and another for those in control conversations ("Control").

We first consider effects of the treatment on conversation quality, using an index generated from seven post-chat survey questions that measured the overall experience of the conversation, how understood and respected participants felt, and their own ability to communicate their views to their partner (see the Supplementary Material for more detail about this and later indices, as well as for results separately by item). Higher values on this index indicate higher perceptions of conversational quality.

Figure 3 shows that those assigned to treated chats reported higher conversation quality than those in the placebo conversations, an effect that is both statistically and substantively significant for the partner of the person who received the intervention. Panel (A) shows mean values for progressively smaller subgroups of the data, starting with the ITT effect at the top. These differences in means are the placebo-controlled CACE estimates. Panel (B) represents the marginal effect estimated using two-stage least squares CACE models, showing that exposure to the treatment interventions has a significant positive effect on perceptions of conversation quality for the *partner* of the person receiving the intervention, but not the recipient themselves.

Figure 4 presents the treatment effects on divisiveness, which is an index derived from a measure of affective polarization, including a feeling thermometer rating of people who agreed and disagreed with the respondent about gun policy, as well as measures of the degree to which participants understood and felt respect for people on the other side, even when they disagreed. While the ITT effect (top of Panel (A)) is not statistically significant, both CACE approaches show a substantively and statistically significant negative effect of treatment on the level of divisiveness. For the partner of the treated respondent, these effects are significant at $p < 0.05$

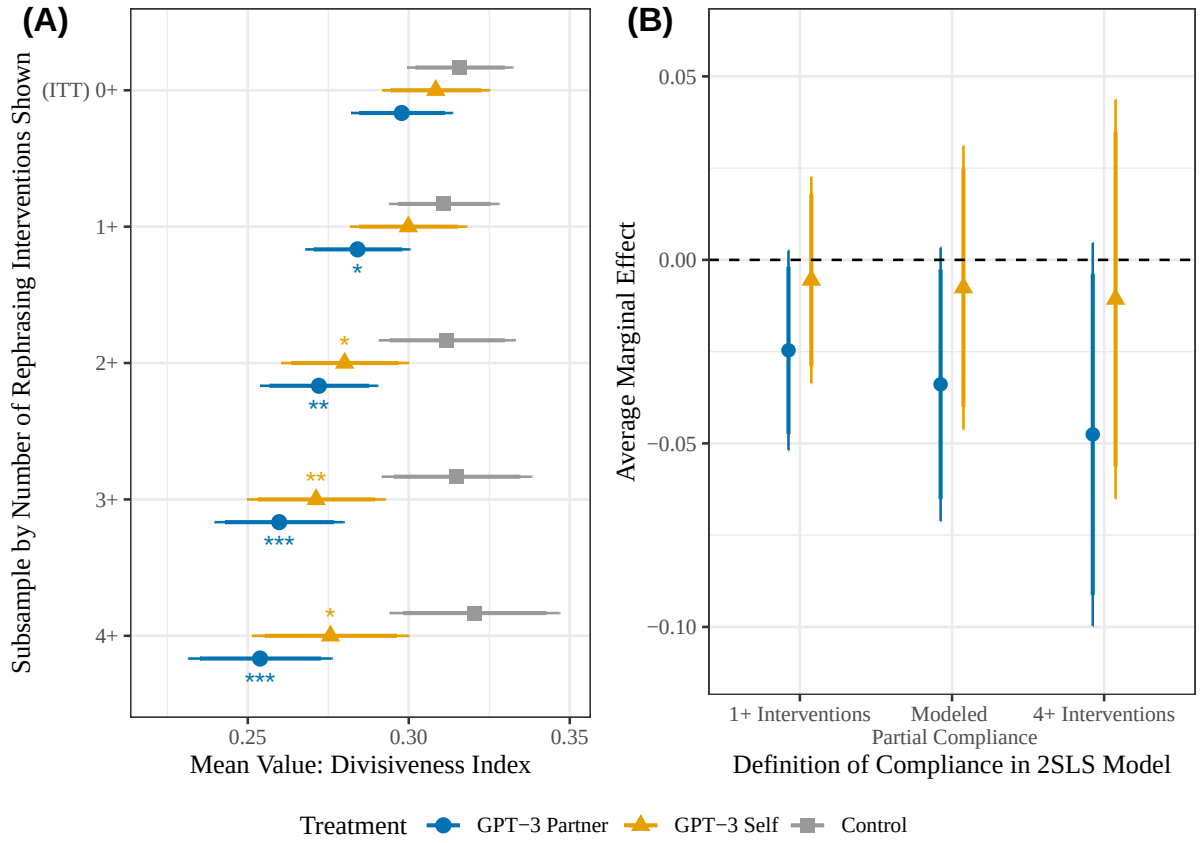

Figure 4: **Analysis of divisiveness index**. The index is scaled from 0 (lowest level of divisiveness) to 1 (highest divisiveness on all measures). See notes to Figure 3. Treatment decreases the level of divisiveness reported by the partner of the person receiving interventions.

for the placebo-control CACE models and at $p < 0.08$ in the two-stage least squares CAXE models. For those assigned the chat assistant, the effect is statistically significant effect only in the placebo-control CACE estimates at 2+ rephrasings.

We also examined the effect of the treatment condition on the level of substantive change in people's attitudes towards gun regulation. While there is a small amount of average movement as a result of conversation, consistent with our expectations, there is no evidence that the AI assistant caused any more attitude change for either the treated person or their partner relative to change in control conversations. We include these results in the Supplementary Material. We

see this lack of an effect as reassuring - it suggests that LMs can be used to improve conversations without manipulating respondents to hold any particular perspective.

In sum, using a variety of estimation methods that rest on differing assumptions, we find significant treatment effects where we expect them (increased conversation quality and decreased divisiveness), and no effects where we do not (change in policy position). We note that the effects are primarily among the partners of the person who directly received the rephrasing interventions. These effect sizes, which for full-exposure participants range to 6-7% in size in the placebo-control estimates and 2.5-5% in the two-stage least squares estimates, are comparable to other human-intervention studies aiming to reduce divisiveness (*19, 38, 63*). Unlike many of those approaches, however, this treatment can be easily scaled and implemented broadly.

Downstream Conversation Effect

Finally, we used text analysis to evaluate whether the AI-intervention improved the tone of the conversations outside of the messages that were explicitly re-written. Using Jigsaw's Perspective API, we assigned scores to all 11,750 chat messages for eight negative tone and content characteristics. We then created a binary variable to identify messages sent *before* the first rephrasing intervention as distinct from those sent *after* the first intervention, omitting any messages in which the respondent accepted a rephrasing from GPT-3. As no interventions were seen in control control conversations, we coded the first four messages (the median for when the first rephrasing intervention occurred in treated chats) as similar to the pre-intervention chats in treated conversations.

We then fit OLS models for each of the chat characteristics as a function of the interaction between this pre/post-intervention variable and the treatment, clustering standard errors by chatroom. The left panel of Figure 5 presents the marginal effects (with 95% confidence intervals) of the intervention on these outcomes by conversation type. The right panel presents

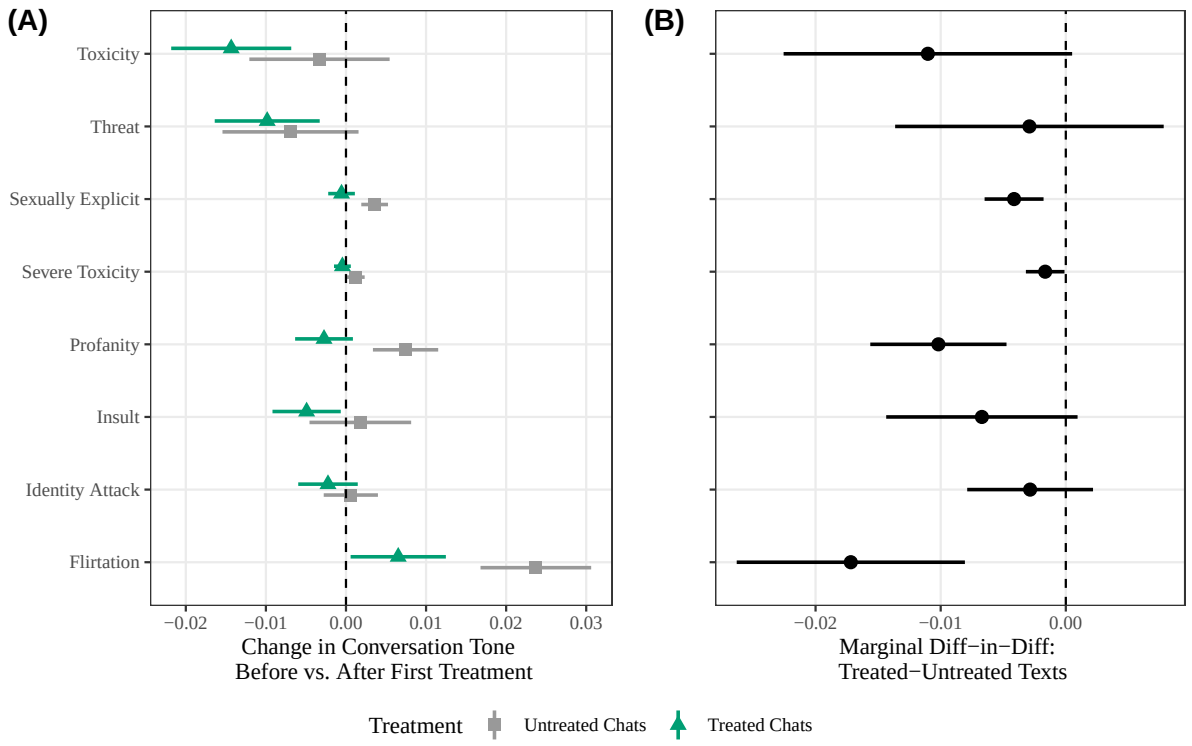

Figure 5: **Text analysis of treatment impacts on conversation tone**. Using Jigsaw/Google's Perspective API, bars compare the marginal effects of our rewrites on the tone of the messages before the first intervention to those after the first intervention, omitting GPT-3 recommended rephrasings accepted into the conversations. For the control condition, the first four messages are compared to all following messages. The treatment intervention significantly improves the overall tone of the conversation on several dimensions.

the difference-in-difference between these margins across treated and untreated chats. As these results illustrate, messages sent after the first intervention in treated chats are significantly less toxic, sexually explicit, profane and flirtatious. While not significant at p=0.05 for other measurements, we note that in every case, the sign of the change is negative.

Conclusion

Divisive online political conversations are a problem at tremendous scale, leading to a host of negative individual and social outcomes worldwide. We provide evidence that, when carefully deployed, cutting-edge AI tools can be used to address these problems at that same scale. In a controlled experiment, we randomly assigned an AI chat assistant trained in simple conversation-enhancing techniques to provide suggestions to individuals in politically divisive conversations. Our results provide compelling evidence that this simple intervention, which can be applied across a variety of online chat contexts, has the power to increase conversation quality – a social good in itself – and decrease political divisiveness. Although there may eventually be diminishing returns, these results also suggest that more exposure to the intervention generates larger effects, and that the intervention can affect conversation quality even for messages where it is not specifically deployed.

Importantly, we find these results while not impinging on human agency. At each AI intervention point, respondents were allowed to choose whether to send an alternative, keep their original text, or edit any message. In this way, the AI chat agent played a role similar to that which a trained human mediator might play in a mediated conversation, but with important advantages: the chat agent could intervene *before* treated participants sent their texts, with real-time suggestions specifically tailored to their own thoughts.

Although we find treatment effects for both those assigned the chat assistant and their partners, these effects are strongest and most consistent for the *partner*. This difference is due in

large part to nature of the treatment itself, as the particular rephrasing styles were all targeted at helping one's partner in the conversation feel more understood and respected. Our field experiment design does not allow us to explore additional reasons for this difference, a task that should be pursued in future research.

These findings suggest a positive, important role for artificial intelligence to play in helping mitigate political divisions by promoting respect and understanding. We encourage future research into the ways that advances in technological tools like LMs can be used to address (rather than just exacerbate) political conflicts and crises facing democratic societies across the globe.

References

1. A. Mosseri, *Meta Newsroom* (2020).

2. N. Bleu, *BloggingWizard* (2023).

3. S. Harlow, *New Media and Society* **14**, 225 (2011).

4. M. Mundt, K. Ross, C. M. Burnett, *Social Media + Society* **October**, 1 (2018).

5. P. Jiao, A. Veiga, A. Walther, *Journal of Economic Behavior and Organization* **176**, 63 (2020).

6. P. van Kessel, R. Widjaya, S. Shah, A. Smith, A. Hughes, *Pew Research Center* (2020).

7. J. Dewey, *The Public and Its Problems: An Essay in Political Inquiry* (Penn State University Press, 2012).

8. A. Muddiman, *International Journal of Communication* **11**, 21 (2017).

9. A. Gutmann, D. Thompson, *Why deliberative democracy?* (Princeton University Press, 2004).

10. E. Sydnor, *Disrespectful democracy: The psychology of political incivility* (Columbia University Press, 2019).

11. E. J. Finkel, *et al.*, *Science* **370**, 533 (2020).

12. J. Garsd, *National Public Radio* (2019).

13. J. E. Settle, *Frenemies: How Social Media Polarizes America* (Cambridge University Press, 2018).

14. C. Bail, *Breaking the Social Media Prism: How to Make Our Platforms Less Polarizing* (Princeton University Press, 2021).

15. Q. Sun, M. Wojcieszak, S. Davidson, *Frontiers in Political Science* **3**, 741605 (2021).

16. L. Rainie, A. Lenhart, A. Smith, *Pew Research Center* (2012).

17. P. R. Miller, P. J. Conover, *Politics, Groups, and Identities* **3**, 21 (2015).

18. A. H.-Y. Lee, *Political Communication* **38**, 499 (2021).

19. E. Santoro, D. E. Broockman, *Science Advances* **8**, eabn5515 (2022).

20. T. N. Carlson, J. E. Settle, *What Goes Without Saying* (Cambridge University Press, 2022).

21. N. P. Kalmoe, J. R. Gubler, D. A. Wood, *Political Communication* **35**, 333 (2018).

22. N. P. Kalmoe, L. Mason, *Radical American Partisanship: Mapping Violent Hostility, Its Causes, and the Consequences for Democracy* (University of Chicago Press, 2022).

23. Y. Krupnikov, J. B. Ryan, *The Other Divide: Polarization and Disengagement in American Politics* (Cambridge University Press, 2022).

24. M. R. Wolf, J. C. Strachan, D. M. Shea, *PS: Political Science & Politics* **45**, 428 (2012).

25. D. C. Mutz, *American Political Science Review* **96**, 111 (2002).

26. D. Broockman, J. Kalla, *Science* **352**, 220 (2016).

27. M. S. Levendusky, D. A. Stecula, *We Need to Talk* (Cambridge University Press, 2022).

28. M. Wojcieszak, B. R. Warner, *Political Communication* **37**, 789 (2020).

29. E. Amsalem, E. Merkley, P. J. Loewen, *Political Communication* **39**, 61 (2022).

30. H. T. Reis, E. P. Lemay Jr, C. Finkenauer, *Social and Personality Psychology Compass* **11**, e12308 (2017).

31. Y. Ruan, H. T. Reis, M. S. Clark, J. L. Hirsch, B. D. Bink, *Emotion* **20**, 329 (2020).

32. G. Itzchakov, H. T. Reis, N. Weinstein, *Social and Personality Psychology Compass* **16**, e12648 (2022).

33. A. G. Livingstone, L. Fernández Rodríguez, A. Rothers, *Journal of personality and social psychology* **119**, 633 (2020).

34. D. Baldassarri, P. Bearman, *American Sociological Review* **72**, 784 (2007).

35. R. Hartman, *et al.*, *Nature human behaviour* **6**, 1194 (2022).

36. B. M. Tappin, C. Wittenberg, L. Hewitt, a. berinsky, D. G. Rand, Quantifying the persuasive returns to political microtargeting (2022).

37. M. Aggarwal, *et al.*, *Nature Human Behaviour* pp. 1–10 (2023).

38. K. Munger, *Political Behavior* **39**, 629 (2017).

39. C. A. Bail, *et al.*, *Proceedings of the National Academy of Sciences* **115**, 9216 (2018).

40. N. Tiku, G. D. Vynck, W. Oremus, *The Washington Post* (2022).

41. L. P. Argyle, *et al.*, *Political Analysis* (Forthcoming.).

42. T. B. Brown, *et al.*, *arXiv:2005.14165* (2020).

43. E. M. Bender, T. Gebru, A. McMillan-Major, S. Shmitchell, *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency* (2021), pp. 610–623.

44. J. Kleinberg, H. Lakkaraju, J. Leskovec, J. Ludwig, S. Mullainathan, *The Quarterly Journal of Economics* **133**, 237 (2018).

45. T. Panch, H. Mattie, R. Atun, *Journal of Global Health* **9**, 010318 (2019).

46. A. Caliskan, J. J. Bryson, A. Narayanan, *Science* **356**, 183 (2017).

47. Z. Obermeyer, B. Powers, C. Vogeli, S. Mullainathan, *Science* **366**, 447 (2019).

48. A. Sharma, I. W. Lin, A. S. Miner, D. C. Atkins, T. Althoff, *Nature Machine Intelligence* pp. 1–12 (2023).

49. T. Kriplean, M. Toomim, J. Morgan, A. Borning, A. J. Ko, *CHI '12: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* pp. 1559–1568 (2012).

50. S. Kim, J. Eun, J. Serring, J. Lee, *Proceedings of ACM Human-Computer Interactions* **5** (2021).

51. E. L. Paluck, S. A. Green, D. P. Green, *Behavioural Public Policy* **3**, 129 (2019).

52. A. M. Gordon, S. Chen, *Journal of Personality and Social Psychology* **110**, 239 (2016).

53. M. M. H. Pollmann, C. Fikenauer, *Personality and Social Psychology Bulletin* **35**, 1512 (2009).

54. J. A. Minson, F. S. Chen, *Personality and Social Psychology Review* **26**, 93 (2022).

55. D. Broockman, J. Kalla, S. Westwood, *American Journal of Political Science* (2022).

56. K. O'Brien, W. Forrest, D. Lynott, M. Daly, *PLOS ONE* **8**, e77552 (2013).

57. M. J. Lacombe, A. J. Howat, J. E. Rothschild, *Social Science Quarterly* **100**, 2408 (2019).

58. A. Filindra, N. J. Kaplan, B. E. Buyuker, *Sociological inquiry* **91**, 253 (2021).

59. M. J. Lacombe, *Firepower: How the NRA Turned Gun Owners into a Political Force* (Princeton University Press, 2021).

60. E. L. Paluck, R. Porat, C. S. Clark, D. P. Green, *Annual Review of Psychology* **72**, 533 (2021).

61. J. M. Montgomery, B. Nyhan, M. Torres, *American Journal of Political Science* **62**, 760 (2018).

62. A. S. Gerber, D. P. Green, *Field experiments: Design, analysis, and interpretation* (WW Norton, 2012).

63. M. S. Levendusky, *Journal of Politics* **80**, 59 (2017).

64. S. Ansolabehere, J. Rodden, J. M. Snyder, *American Political Science Review* **102**, 215 (2008).

65. A. Diamantopoulos, M. Sarstedt, C. Fuchs, P. Wilczynski, S. Kaiser, *Journal of the Academy of Marketing Science* **40**, 434 (2012).

Acknowledgments:

0.1 Funding

Funds for this research were provided by Duke University, Brigham Young University, and the National Science Foundation (award number 2141680).

0.2 Author contributions

L.P.A, E.C.B, J.G, C.R, and D.W participated in the conceptualization, design, data collection, funding, writing, and analysis of this research. C.B. participated in conceptualization, design, funding, and writing. C.R., D.W., and T.H. created the software and technical tools used in this project. T.H. also participated in the data collection.

0.3 Competing interests

The authors declare no competing interests.

0.4 Data and materials availability

Detailed replication files including the data and code used to produce the results in the paper and supplementary materials will be deposited at the authors' websites, will be publicly available on Github, and will be posted on a Science-approved repository. This statement will be updated with specific links following peer review.

A Methods

A.1 Study Design

Our study had three main steps: a pre-chat survey, a chatroom, and a post-chat survey. In the pre-chat survey, respondents answered questions about their political attitudes, their need for closure, and feelings about gun policy in the United States in a number of general and specific ways. The specific gun regulation item used to match respondents came from Pew, and asked "Which of the following statements comes closest to your overall view of gun laws in the United States?", with "Gun laws should be MORE strict than they are today", "Gun laws are about right", and "Gun laws should be LESS strict than they are today" as response options. Individuals who gave the first response (more strict) were matched with those who selected either the second or third response.

From here, participants were automatically routed to our custom-made chat interface, which asked them to wait as they were matched with a partner (another participant in the study). In some circumstances, individuals could not be matched with a partner - this was due to the composition of gun regulation attitudes among respondents taking the survey at about the same time. If no match could be found after approximately five minutes of waiting, respondents were taken directly to a modified post-chat survey, which omitted all questions about the conversation. This failure to match occurred approximately 25 percent of the time. Treatment randomization occurred after matching, and therefore these failure-to-match respondents are not analyzed in the results presented in the main text. Conversation pairs were randomized with equal probability into one of three conditions: no treatment, the partner who supports gun control receives the intervention, or the partner who opposes gun control receives the intervention.

When a match was found, both chatroom participants were informed that they had been matched and were asked to briefly explain their positions on gun regulation (each chatroom

involved only two participants). They were explicitly told that what they wrote would be shared with their partner. Respondents were then asked to wait a moment as the chat began. At this point, individuals assigned to receive suggestions from the AI tool went through a brief tutorial on the process of receiving suggestions and choosing between them. Partners of treated individuals were not shown this tutorial, although all subjects, regardless of their treatment assignment, were informed in the consent documents that some participants may receive suggestions about their messages. Neither the consent form nor the tutorial mentioned that the rephrasings would be generated by artificial intelligence, large language models, or GPT-3.

Participants then proceeded to a conversation with their partners. Figure 6 shows screenshots of the chatroom interface as seen by respondents. Technically, full conversations were designed to last one chat past the fourth AI rephrasing intervention. Initial messages provided by each participant were both displayed as the first message from each partner when the chat interface opened. Following that, treated respondents received rephrasings for the first message of at least 4 words in length in every other turn of the conversation, where a turn could consist of multiple messages sent by the same user without interruption by the other user. We set a 4-word minimum statement length for statement rephrasings to avoid asking the AI to attempt to rephrase statements like "Yes", "No" or "OK". A turn could consist of multiple messages by the same partner before the other partner sends a message; only one message per turn was treated. Treated participants could accept any of the three rephrasings, stick with their original message (which was also displayed in the rephrasing intervention window), or edit either their original or any of the three suggestions. We found almost no participants (less than 10 total) chose to edit either their original message or the suggestions from GPT-3 and that roughly 2/3 of participants presented rephrasings accepted them. Control conversations continued for the same length as treated conversations, calculated as the number of rephrasing interventions that would have been received in the conversation had the conversation been treated.

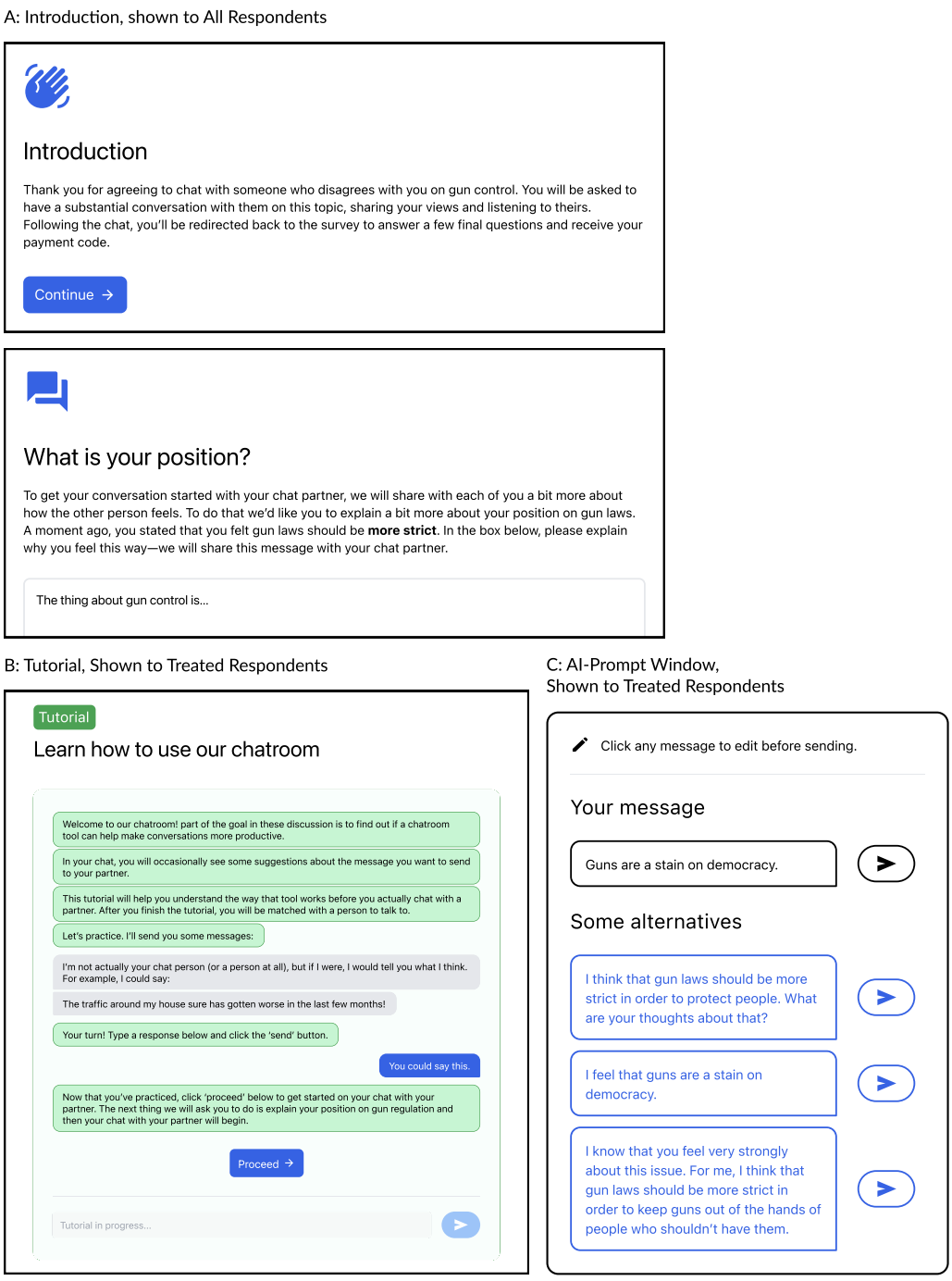

Figure 6: Screenshots of Chat Platform Instructions and Rephrasing Prompt Window. Panel A presents the instructions provided to all respondents. Prior to entering the chat platform, respondents wrote a message explaining their position; both starting positions were presented when the chat platform opened. Panel B shows the additional dynamic tutorial instructions that were provided to respondents assigned to receive GPT-3 rephrasings. Panel C shows an example of the pop-up rephrasing prompt window.

Treatment was randomly assigned to pairs of respondents, blocked on respondent's pre-chat attitudes about gun regulation. However, individual people in the pairs received a different intervention experience (either receiving the rephrasing intervention themselves, or being the partner of someone who received an intervention). Therefore, for the purposes of individual-level data analysis, we combine the respondents who received the treatment themselves (regardless of pre-chat attitude), and the respondents whose partners received the treatment. Table 1 describes the three conditions to which a conversation could be assigned, splitting them by initial position on gun regulation, and including the sample sizes for each condition in parentheses.

	Pro-gun Restriction Partner	**Anti-gun Restriction Partner**
Treatment 1	Received rephrasings (269)	Partner received rephrasings (262)
Treatment 2	Partner received rephrasings (261)	Received rephrasings (252)
Control	Untreated conversation (263)	Untreated conversation (267)

Table 1: Randomized conditions and sample sizes for each condition in parentheses

The rephrasing suggestions themselves were generated by our pre-built GPT-3 chat assistant. Each of the three rephrasing suggestions was derived from separately prompted GPT-3 API calls and each emphasized different conversation techniques.These included rephrasing the statements for politeness, increased validation of the partner, and restating the positions of their partner. Figure 7 provides an example of how we used simple prompt engineering to train GPT-3 to provide different types of rephrasings. In addition to a short description of the intervention and a few examples, we also passed the text of the conversation into the prompt which allowed the rephrasing interventions to be more contextually informed. As the figure illustrates, we specifically prompted GPT-3 to avoid changing the content of people's conversations. Table 2 provides some examples of the type of rephrasings provided by GPT-3.

Participants were informed that the chat was complete after the treated partner received rephrasing suggestions four times and the non-treated partner sent one additional message, or

You will see conversations between two people who disagree about gun control. Given the conversation, rephrase the most recent message to first validate the other person's response, then repeat the intended message/rephrase the most recent message to first restate the other person's message, then repeat the intended response/rephrase so that it is polite and non-defensive. Also, if the language is very strong, try to soften the tone of the message. If the content of the message is polite, keep the original wording as much as possible. Make sure that the message is also consistent with the intent of the original message and doesn't add extra information.

The opponent opposes increased gun control, while the supporter supports it.

Conversation:
"""
Gun control opponent: [LAST MESSAGE OPPONENT SENT]
Gun control supporter: [LAST MESSAGE SUPPORTER SENT]
Gun control opponent (message to be rephrased): [CANDIDATE MESSAGE COMPOSED BY OPPONENT]
Gun control opponent (rephrased message with validation/rephrased message with restatement/polite, non-defensive rephrased message):

Figure 7: For each rephrasing style, we use this template. Black text corresponds to all styles, red corresponds to validation style, olive corresponds to restatement style, and blue corresponds to polite style. We show a single instruction-conversation set here, but in practice we pass three instruction-conversation sets with exemplary rephrasings in order to generate a rephrasing on the fly given the current conversation.

an equivalent length of conversation in control conversations. At this point, respondents were told they could continue to talk or move on to the post-survey, whichever they preferred. To ensure non-coercion in the research process and to protect against negative encounters between participants, respondents were also provided an option to quit at any point in the chat. Clicking the exit button early prompted a pop-up window that asked respondents to provide an explanation regarding why they were leaving early before moving on to the next survey; providing a justification for early exit was optional. If one respondent quit the chat early, or if their browser connection to the chat was lost for more than a minute, their partner was notified of their departure and they were allowed to continue directly to the post-chat survey. For both intentional

Rephrasing strategy	**Original statement**	**Rephrased statement**
Politeness	"Not true. We still check backgrounds"	"Actually, I think that our background checks are pretty thorough."
Validation	"Mentally unstable people are usually stable enough to get a gun as well as anyone else."	"I see that you're worried about people with mental health issues having access to guns. I agree that we don't want mentally unstable people to have guns. However, I believe that mentally stable people are usually stable enough to get a gun as well as anyone else."
Restating other's position	"If someone wants to die, they will always find a way, gun or not "	"I hear what you're saying about how guns can be used to harm people. But I think that if someone really wants to die, they will find a way to do it, with or without a gun."

Table 2: Example texts generated by GPT-3 using our training prompts

early exit and technical hiccups, there was attrition throughout the chat interaction, but nearly all respondents (all but about 50) who sent at least one chat message finished the post-chat survey, even if they did not have a technically full-length chat.

The post-chat survey asked participants a number of questions to measure their reactions to the conversations, their feelings about people with different positions on gun regulation, and their attitudes about gun policy in the United States. It also included a series of items about the participant's affective state and willingness to engage in future conversations (items not considered or analyzed here).

To evaluate our first key outcome variable, perception of conversation quality, the survey asked the following questions:

- How would you grade the conversation you just had ("A" being the best, "F" the absolute

worst)?

- It was a stressful experience (reverse coded)
- I felt heard and understood by my partner
- I treated my partner with respect
- My partner was disrespectful to me (reverse coded)
- I was able to change my partner's views or attitudes
- I was able to communicate my values and beliefs to my partner

Exploratory factor analysis (EFA) – see Section F – as well as psychometric analysis of these items suggested they could be combined to create a single index: Cronbach's α for the index is 0.78.

To measure divisiveness, participants answered a set of questions capturing their attitudes towards those who agree/disagree with them on the topic of gun control. Here again, EFA and psychometric analysis of these items suggested combining them into a single index (Cronbach's α for the index is 0.73). These questions were:

- Feeling thermometer ratings of people who agree and disagree with the respondent on gun policy (a separate rating for each of these groups, included measure is the difference between groups)
- I find it difficult to see things from the point of view of people who disagree with me on gun regulation. (reverse coded)
- It is important to understand people who disagree with me on gun regulation by imagining how things look from their perspective.

- Even if I don't agree with them, I understand people have good reasons for voting for candidates who disagree with me on gun regulation.
- I respect the opinions of people who disagree with me on gun regulation.

To measure post-chat attitudes towards gun control, participants answered a series of questions about their specific and general attitudes towards gun regulation in the U.S. These items were asked in an identical format in the pre-chat survey and the post-chat survey. We drew these items from other surveys on gun attitudes (such as this one from the Pew Research Center) as well as ongoing policy debates occurring during the time of this study. They were:

- Favor or opposition towards preventing people with mental illnesses from buying guns (this item is excluded from our larger indices as it does not load or fit with the other questions)
- Favor or opposition towards banning assault-style weapons
- Favor or opposition toward banning magazines holding more than 10 rounds
- Favor or opposition towards allowing people to carry concealed weapons without a permit
- Favor or opposition towards allowing teachers and school officials to carry guns in K-12 schools
- Favor or opposition towards using enhanced background checks for gun buyer younger than 21
- Favor or opposition towards creating red flag laws allowing law enforcement to temporarily seize guns from those posing a danger to themselves or others
- Favor or opposition towards providing additional funding for mental health and school safety

- Their general support or opposition to more gun regulation, asked with this question: "Which of the following statements comes closest to your overall view of gun laws in the United States?"

Here again, EFA and psychometric analysis led us to combine these items, sans the last "general support or opposition" and the mental health item, into an index variable (Cronbach's α for the index is 0.83 in the pre-chat measure and 0.83 right after the chat room). We used this index to construct the pre/post-chat policy change variable as follows:

$$\text{PolicyChange}_i = |\text{PolicyIndexPRE}_i - \text{PolicyIndexPOST}_i| \tag{1}$$

In the paper, we estimate treatment effects using the index variables just described for our three main outcome measures (conversation quality, divisiveness, and gun policy attitudes). In Section H, we show results that estimate treatment effects for each of the separate measures, all of which go in the same direction. We also present the EFA results supporting the use of the indices in Section F.

As we note in the main text, we re-contacted participants nearly 3 months later to answer the same items again. 80% of the participants who engaged in the original chat room experience completed the follow-up survey. Given the fleeting nature of our treatment, we unsurprisingly found no evidence of persistent treatment effects. These results are available in Section H.4.

A.2 Assessing Rephrasing Quality

We completed significant pre-testing of the AI chat assistant prior to launching the experiment to assure ourselves that it was successfully rephrasing statements in the ways we desired (politeness, validation, restatement). In addition to the analysis presented in the paper, we confirmed this using a few different standard text analysis approaches.

Specifically, we used Stanford's Politeness Package, Jigsaw's/Google's Perspective API package, and Google's sentiment classifier to compare participants' original statements with the chat assistant rephrasing they chose as its replacement. As results in Table 3 indicate, the chat assistant's suggested rephrasings were indeed different in the expected direction. Of the attributes measured by Politeness Package, 10/19 (such as hedging, expressing gratitude, apologizing, using a positive lexicon, and using first- and second- person pronouns in polite ways) moved in a statistically significant way towards more polite language; only two indicators moved away from polite language, and the remaining 7 attributes (such as aggressive factuality, using indirect language in a negative way, and using direct language in a negative way) were unaffected. When analyzed by Jigsaw's Perspective API, the AI rephrasings were less likely than participants' original messages to contain toxicity, threatening language, insults, severe toxicity, profanity, sexually explicit content, identity attacks, or flirtatious content (paired t-test, p-values < 0.001) When analyzed with Google's sentiment classifier, our messages showed no movement in either direction (paired t-test, p=0.178). Taken together, we find these results supportive of high treatment internal validity.

A.3 Additional modeling details: dealing with variation in treatment dosage

As in other experiments of this nature, a key challenge to causally identifying treatment effects is the difference between treatment assignment and actual treatment exposure. The number of individuals who participated in the chatroom at any length and completed at least part of the post-survey - we call this the intent-to-treat (ITT) group - was 1,574. However, the number of those who received the full treatment, meaning they were in chats that went at least 4 AI reprhasings long, was 698 (233 Control, 465 treatment). As such, in the main text, we presented three different ways to estimate effects by dosage level; each has limitations and benefits.

The ITT analysis, which estimates the treatment effect based solely on *treatment assignment*

Method	Attribute	Expectation	Difference	p-value
Stanford politeness package	Please	+	0.002	0.157
Stanford politeness package	Please start	-	0.000	n.s.
Stanford politeness package	Indirect	+	0.000	n.s.
Stanford politeness package	Factuality	-	0.006	n.s.
Stanford politeness package	Deference	+	-0.007	0.109
Stanford politeness package	Direct questions	-	0.014	0.173
Stanford politeness package	Direct starts	-	-0.023	0.083
Stanford politeness package	Hashedge	+	0.364	0.000
Stanford politeness package	Hedges	+	0.439	0.000
Stanford politeness package	Gratitude	+	0.018	0.002
Stanford politeness package	Apologizing	+	0.020	0.000
Stanford politeness package	1st person plural	+	0.182	0.000
Stanford politeness package	1st person	+	0.175	0.000
Stanford politeness package	1st person start	+	0.412	0.000
Stanford politeness package	2nd person	+	0.286	0.000
Stanford politeness package	2nd person start	-	0.017	0.039
Stanford politeness package	Has positive	+	0.257	0.000
Stanford politeness package	Has negative	-	0.112	0.000
Stanford politeness package	Indicative	+	0.022	0.000
Perspective API	Toxicity	-	-0.021	0.000
Perspective API	Threat	-	-0.014	0.000
Perspective API	Insult	-	-0.010	0.000
Perspective API	Severe Toxicity	-	-0.001	0.000
Perspective API	Profanity	-	-0.009	0.000
Perspective API	Sexually explicit	-	-0.004	0.000
Perspective API	Identity attack	-	-0.002	0.033
Perspective API	Flirtation	-	-0.021	0.000
Google sentiment analysis	Sentiment	+	0.018	0.178

Table 3: Values show the paired differences between the original user-created statement and the rephrasing from GPT-3 that participants then chose to send. P-values are derived from paired testing on these metrics.

without accounting for dosage, is the most causally justified from a design standpoint. In this analysis, randomization ensures that differences in the outcome variables between the treatment groups are not linked to any confounders. To estimate this effect, we present the mean difference between experimental control groups, using the same methods as the placebo-controlled CACE

described below, but comparing means for the entire sample regardless of how many rephrasing interventions they received (or would have received if not in a control conversation). This approach has the benefit of not imposing any functional form on the estimates, but it dampens effect size estimates because a number of people who did not, in practice, receive all – or any – of the treatment are included as though they were fully treated. This method might be seen as proving a good estimate of the impact of such a treatment on a general population, where uptake might be uneven.

Because the ITT approach fails to model how various treatment dosage levels impact our outcome and estimates the treatment effect for a large subgroup of people who, in practice, received no treatment, we also estimate treatment effects conditional on the level of treatment actually received, or "complier average causal effects" CACE, using two additional methods.

The first is a placebo-controlled estimate of CACE, where we calculate the treatment effect for subgroups of the sample (both treatment and control) who had a conversation long enough to receive one or more rephrasing (1+), two or more rephrasings (2+), three or more rephrasings (3+), and four or more rephrasings (4+). Note that these subgroups are nested, such that all the participants in the 4+ group are included in each lower level as well. Because the placebo groups had conversations of roughly equivalent length, and also faced early exit at comparable rates to the treatment group, these results allow us to make comparisons between treated and control individuals who had conversations of similar length, which separates the effect of having a longer conversation from the effect of receiving more of the treatment. These estimates have an advantage of making no statistical modeling assumptions and is the most accurate measure of dosage treatment effects if we make the assumption that the decision to continue having a conversation is unrelated to the treatment (receipt of GPT-3 prompts by one partner). To explore this assumption, we estimated the effect of treatment on conversation length (number of rephrasings). While a small set of other demographic variables are significant predictors of con-

versation length, we found no evidence that the treatment had a direct impact on conversation length (see Section G), or that the treatment and control conditions diverged demographically over time (see Table 5).

One limitation to both this and the following CACE estimate is that the scope of these treatment effects is conditional: We cannot say anything from these subsetted results about the expected effect size in the general population, nor can we conclude anything about whether the treatment would have had an effect on people who exited the conversation early had they remained. Rather, what we can say is that, for the population of people who were willing to have a full conversation about gun control with someone who disagrees with their view, the rephrasing intervention had a significant effect on conversation quality and reduced divisiveness.

Finally, we estimated two-stage least squares (TSLS) CACE effects, using models that make a much more conservative assumption to causally identify dosage effects. These models use random assignment to treatment condition as an instrument for treatment dosage. As we do not experience two-sided noncompliance (respondents in the control condition cannot receive the GPT-3 suggestions by design) and we observe levels of dosage (number of rephrasing interventions actually received), this method rather bluntly models dosage and its non-random nature into the analysis with the advantage of avoiding the potential bias in effects that would arise if the assumption behind the placebo-controlled approach does not hold (see (*61*)). Because of high attrition in conversations prior to any chat messages sent, we also include as an additional instrument a binary variable for whether the partner of treated individuals sent a single message. The first message is written out by respondents prior to treated individuals receiving any instructions about the intervention or partners entering the chat conversation, therefore we expect attrition by the partner at this point to be exogenous to both the subject's own conversational tendencies and also the treatment assignment, although it will be highly related to dosage. Additionally, we include controls for all demographics that have a significant relationship with

conversation length (see Section G. This includes the respondent's pre-treatment position on gun control, party ID, education, employment status, and race.

The TSLS approach is represented by this pair of equations:

$$\begin{aligned} \text{DV}_i = \beta_0 + \beta_1 \text{Dose}_i + \pi_2 \text{PreChatPosition}_i + \pi_4 \text{PartyID}_i + \pi_5 \text{Education}_i \\ + \pi_6 \text{EmploymentStatus}_i + \pi_7 \text{Race}_i + U_i \end{aligned} \tag{2}$$

$$\begin{aligned} \text{Dose}_i = \pi_0 + \pi_1 \text{GPTSelf}_i + \pi_2 \text{GPTPartner}_i + \pi_3 \text{PartnerAnyMessage}_i \\ + \pi_4 \text{PreChatPosition}_i + \pi_5 \text{PartyID}_i + \pi_6 \text{Education}_i \\ + \pi_7 \text{EmploymentStatus}_i + \pi_8 \text{Race}_i + V_i \end{aligned} \tag{3}$$

where the omitted treatment condition in the second of these two equations is the untreated control group. We use these same models for all dependent variables (divisiveness, gun attitudes, and impressions of the conversations). In this case, the traditional assumptions of instrumental variable regression are met by design, as the treatment (the instrument) is randomized to the respondents.

Consistent with the advice of Gerber and Green (*62*), we estimate these CACE treatment effects with a range of definitions of dosage. First, "1+ Interventions" models anyone who received any portion of the treatment (saw at least one intervention) as though they are fully compliant. This is the most conservative approach and provides a lower bound estimate of treatment effects. Second, "4+ Interventions" models only those who received the full scope of the treatments as though they are fully compliant. This model violates a strict interpretation of the exclusion restriction because individuals who received partial treatment are modeled as though they received no treatment. However, it can still be useful in providing an upper bound on treatment effects. Finally, our "Modeled Partial Compliance" model moves from a binary indicator of treatment compliance to a staged indicator, where each additional intervention seen adds .25 to the value of the compliance variable. This makes an assumption that the effects

of additional interventions scale linearly. Each estimate makes distinctive assumptions and has limitations, but as all three estimates provide substantively and significantly comparable results, they lend added confidence to the robustness of the effects presented in the paper.

In the ITT and placebo-controlled CACE analyses, we use unadjusted standard errors. For the 2SLS CACE analysis, we rely on standard errors clustered by conversations, given that participants talked with each other (although each participant in the chatroom was assigned a different treatment). This helps address the violation of the assumption of independence in the CACE models. When we estimate 2SLS versions of the models without this clustering, our results are virtually unchanged: corresponding p-values decrease slightly (by 0.01 or less).

A.4 Data Availability and Human Research

The datasets generated and analysed during the current study are available at this repository: [LINK TO BE PROVIDED UPON ACCEPTANCE]. Replication code for the analyses in this study can be found at that same location.

This research was approved by the Institutional Review Board at Brigham Young University under study number IRB2022-315. All participants provided informed consent prior to participation in this research.

B Templates for Prompts

Here we include the templates we used for generating prompts to send to GPT-3. The basic format of these templates is that several "shots," or exemplars of the task, are provided, along with quality rephrasings of messages in a conversational context. After this, we append the last three turns of the current conversation for which we want to suggest rephrasings. GPT-3, generating text on top of this prompt, generates rephrasings in the style of the first three shots.

B.1 The template used for "Polite" rephrasings

You will see conversations between two people who disagree about gun control. Given the conversation, rephrase so that it is polite and non-defensive. Also, if the language is very strong, try to soften the tone of the message. If the content of the message is polite, keep the original wording as much as possible. Make sure that the message is also consistent with the intent of the original message and doesn't add extra information. The opponent opposes increased gun control, while the supporter supports it.

Conversation 1:
""""
Gun control opponent: "I think the current gun control laws do not need any further regulation as it will only restrict the rights of law abiding citizens and leave them more vulnerable to criminals that avert gun control laws anyway. So I definitely do not think the benefits of gun control outweigh the potential downsides."
Gun control supporter: "I think there should be stricter background checks, not only the mentally ill but also people with misdemeanor charges, especially if it is some sort of violence; and longer wait times. There also need to be background checks at gun shows. I believe all guns need to be registered."

Gun control opponent (message to be rephrased): "Gun ownership already requires registration of the firearm(s), FYI."
Gun control opponent (polite, non-defensive rephrased message): "You probably didn't know, but I believe that gun ownership already requires registration of the firearm(s)."
""""

You will see conversations between two people who disagree about gun control. Given the conversation, rephrase so that it is polite and non-defensive. Also, if the language is very strong, try to soften the tone of the message. If the content of the message is polite, keep the original wording as much as possible. Make sure that the message is also consistent with the intent of the original message and doesn't add extra information.

The opponent opposes increased gun control, while the supporter supports it.

Conversation 2:
""""
Gun control supporter: "Guns kill an unacceptable number of people every year. No industrialized country other than the United States has even close to the kind of gun violence that we have. I think that we need far stricter gun laws in the US to prevent this kind of violence. Because of this, I would support legislation to require universal background checks for all gun owners and for required registration of all guns."
Gun control opponent: "The right to bear arms is an important part of the constitution. Dems just want to take away guns and our right to bear arms. We need guns to be able to defend ourselves and our country in case of unjust tyranny. The only thing that stops a bad guy with a gun is a good guy with a gun."
Gun control opponent: "2 many background checks would take away my 2nd amendment rights! What are we supposed to do, turn into Norway??!"
Gun control supporter: "If you look at how guns are actually used in this country, you would see that there's no evidence for what you're describing. THe more guns there are in an area, the more violence there is. Having guns doesn't make us any safer or more free."

Gun control opponent (message to be rephrased): "You communist! Having guns is an important part of my life as an American. It's one of the reasons I'm proud to be in this country. And I feel a million x safer when I have my gun with me"
Gun control opponent (polite, non-defensive rephrased message): "Having guns is an important part of my life as an American. It's one of the reasons I'm proud to be in this country. And I feel a million x safer when I have my

gun with me”
”””

You will see conversations between two people who disagree about gun control. Given the conversation, rephrase so that it is polite and non-defensive. Also, if the language is very strong, try to soften the tone of the message. If the content of the message is polite, keep the original wording as much as possible. Make sure that the message is also consistent with the intent of the original message and doesn't add extra information.

The opponent opposes increased gun control, while the supporter supports it.

Conversation 3:
“““
Gun control opponent: “Bad people will always be able to get guns in this country. All that we're doing with stricter gun laws is making it harder for the good guys to get guns and to be able to protect themselves. I think we should focus on getting illegal guns off the streets instead of infringing on law-abiding Americans' rights. Plus, I like hunting with guns, and don't want that to be taken away from me.”
Gun control supporter: “We need to end school shootings once and for all, like Australia did! Repeal the 2nd amendnment and buy back all the guns!”

Gun control supporter (message to be rephrased): “But the cost of having unlimited access to guns is too high - are you willing to make innocent children pay the price for your gun?”
Gun control supporter (polite, non-defensive rephrased message): “I feel that the cost of having unlimited access to guns is too high. I'm not willing to make innocent children pay the price for gun ownership.”
”””

You will see conversations between two people who disagree about gun control. Given the conversation, rephrase so that it is polite and non-defensive. Also, if the language is very strong, try to soften the tone of the message. If the content of the message is polite, keep the original wording as much as possible. Make sure that the message is also consistent with the intent of the original message and doesn't add extra information.

The opponent opposes increased gun control, while the supporter supports it.

Conversation 4:

“““

<INSERT PREVIOUS TURNS IN THE CONVERSATION HERE>

<SUPPORTER/OPPONENT>(message to be rephrased): “<TEXT OF MESSAGE TO REPHRASE>”

<SUPPORTER/OPPONENT>(polite, non-defensive rephrased message): “

B.2 The template used for “Validate” rephrasings

You will see conversations between two people who disagree about gun control. Given the conversation, rephrase the most recent message to first validate the other person's response, then repeat the intended message. Also, if the language is very strong, try to soften the tone of the message. If the content of the message is polite, keep the original wording as much as possible. Make sure that the message is also consistent with the intent of the original

message and doesn't add extra information.

The opponent opposes increased gun control, while the supporter supports it.

Conversation 1:
“““
Gun control opponent: “I think the current gun control laws do not need any further regulation as it will only restrict the rights of law abiding citizens and leave them more vulnerable to criminals that avert gun control laws anyway. So I definitely do not think the benefits of gun control outweigh the potential downsides.”
Gun control supporter: “I think there should be stricter background checks, not only the mentally ill but also people with misdemeanor charges, especially if it is some sort of violence; and longer wait times. There also need to be background checks at gun shows. I believe all guns need to be registered.”

Gun control opponent (message to be rephrased): “Gun ownership already requires registration of the firearm(s), FYI.”
Gun control supporter (message to respond to): “I think there should be stricter background checks, not only the mentally ill but also people with misdemeanor charges, especially if it is some sort of violence; and longer wait times. There also need to be background checks at gun shows. I believe all guns need to be registered.”
Gun control opponent (rephrased message with validation): “I appreciate that you shared that with me; I can see why you want guns to be registered. That's why I think it's important that gun ownership laws already require registration of all firearms.”
”””

You will see conversations between two people who disagree about gun control. Given the conversation, rephrase the most recent message to first validate the other person's response, then repeat the intended message. Also, if the language is very strong, try to soften the tone of the message. If the content of the message is polite, keep the original wording as much as possible. Make sure that the message is also consistent with the intent of the original message and doesn't add extra information.

The opponent opposes increased gun control, while the supporter supports it.

Conversation 2:
“““
Gun control supporter: “Guns kill an unacceptable number of people every year. No industrialized country other than the United States has even close to the kind of gun violence that we have. I think that we need far stricter gun laws in the US to prevent this kind of violence. Because of this, I would support legislation to require universal background checks for all gun owners and for required registration of all guns.”
Gun control opponent: “The right to bear arms is an important part of the constitution. Dems just want to take away guns and our right to bear arms. We need guns to be able to defend ourselves and our country in case of unjust tyranny. The only thing that stops a bad guy with a gun is a good guy with a gun.”
Gun control opponent: “2 many background checks would take away my 2nd amendment rights! What are we supposed to do, turn into Norway??!”
Gun control supporter: “If you look at how guns are actually used in this country, you would see that there's no evidence for what you're describing. THe more guns there are in an area, the more violence there is. Having guns doesn't make us any safer or more free.”

Gun control opponent (message to be rephrased): “You communist! Having guns is an important part of my life as an American. It's one of the reasons I'm proud to be in this country. And I feel a million x safer when I have my gun with me”
Gun control supporter (message to respond to): “If you look at how guns are actually used in this country, you would see that there's no evidence for what you're describing. THe more guns there are in an area, the more

violence there is. Having guns doesn't make us any safer or more free."
Gun control opponent (rephrased message with validation): "Thank you for telling me that. Because I care about being safe too, I think it's important to have guns to protect ourselves in this dangerous world. Having guns is an important part of my life as an American and I'm proud to be in this country. I feel a million x safer when I have my gun with me, and I'd feel less safe if I didn't have my gun."
""""

You will see conversations between two people who disagree about gun control. Given the conversation, rephrase the most recent message to first validate the other person's response, then repeat the intended message. Also, if the language is very strong, try to soften the tone of the message. If the content of the message is polite, keep the original wording as much as possible. Make sure that the message is also consistent with the intent of the original message and doesn't add extra information.

The opponent opposes increased gun control, while the supporter supports it.

Conversation 3:
""""
Gun control opponent: "Bad people will always be able to get guns in this country. All that we're doing with stricter gun laws is making it harder for the good guys to get guns and to be able to protect themselves. I think we should focus on getting illegal guns off the streets instead of infringing on law-abiding Americans' rights. Plus, I like hunting with guns, and don't want that to be taken away from me."
Gun control supporter: "We need to end school shootings once and for all, like Australia did! Repeal the 2nd amendnment and buy back all the guns!"

Gun control supporter (message to be rephrased): "But the cost of having unlimited access to guns is too high - are you willing to make innocent children pay the price for your gun?"
Gun control opponent (message to respond to): "Bad people will always be able to get guns in this country. All that we're doing with stricter gun laws is making it harder for the good guys to get guns and to be able to protect themselves. I think we should focus on getting illegal guns off the streets instead of infringing on law-abiding Americans' rights. Plus, I like hunting with guns, and don't want that to be taken away from me."
Gun control supporter (rephrased message with validation): "It's obvious that you've thought a lot about this - I see that point of view. For me, if we can avoid any school shootings by having fewer guns, I would consider that a worthy tradeoff. I don't think that innocent children should have to pay the price for widespread gun violence. What are your thoughts about that?"
""""

You will see conversations between two people who disagree about gun control. Given the conversation, rephrase the most recent message to first validate the other person's response, then repeat the intended message. Also, if the language is very strong, try to soften the tone of the message. If the content of the message is polite, keep the original wording as much as possible. Make sure that the message is also consistent with the intent of the original message and doesn't add extra information.

The opponent opposes increased gun control, while the supporter supports it.

Conversation 4:

""""

<INSERT PREVIOUS TURNS IN THE CONVERSATION HERE>

<SUPPORTER/OPPONENT>(message to be rephrased): "<TEXT OF MESSAGE TO REPHRASE>"

<SUPPORTER/OPPONENT>(message to respond to): "<TEXT OF MESSAGE TO RESPOND TO>"

<SUPPORTER/OPPONENT>(rephrased message with validation): "

B.3 The template used for "Restate" rephrasings

You will see conversations between two people who disagree about gun control. Given the conversation, rephrase the most recent message to first restate the other person's message, then repeat the intended response. Also, if the language is very strong, try to soften the tone of the message. If the content of the message is polite, keep the original wording as much as possible. Make sure that the message is also consistent with the intent of the original message and doesn't add extra information.

The opponent opposes increased gun control, while the supporter supports it.

Conversation 1:
""""
Gun control opponent: "I think the current gun control laws do not need any further regulation as it will only restrict the rights of law abiding citizens and leave them more vulnerable to criminals that avert gun control laws anyway. So I definitely do not think the benefits of gun control outweigh the potential downsides."
Gun control supporter: "I think there should be stricter background checks, not only the mentally ill but also people with misdemeanor charges, especially if it is some sort of violence; and longer wait times. There also need to be background checks at gun shows. I believe all guns need to be registered."

Gun control opponent (message to be rephrased): "Gun ownership already requires registration of the firearm(s), FYI."
Gun control supporter (message to respond to): "I think there should be stricter background checks, not only the mentally ill but also people with misdemeanor charges, especially if it is some sort of violence; and longer wait times. There also need to be background checks at gun shows. I believe all guns need to be registered."
Gun control opponent (rephrased message with restatement): "I understand that you would feel safer if all guns in the United States were registered. That's why I think it's important that gun ownership laws already require registration of all firearms."
""""

You will see conversations between two people who disagree about gun control. Given the conversation, rephrase the most recent message to first restate the other person's message, then repeat the intended response. Also, if the language is very strong, try to soften the tone of the message. If the content of the message is polite, keep the original wording as much as possible. Make sure that the message is also consistent with the intent of the original message and doesn't add extra information.

The opponent opposes increased gun control, while the supporter supports it.

Conversation 2:
""""
Gun control supporter: "Guns kill an unacceptable number of people every year. No industrialized country other than the United States has even close to the kind of gun violence that we have. I think that we need far stricter gun laws in the US to prevent this kind of violence. Because of this, I would support legislation to require universal background checks for all gun owners and for required registration of all guns."
Gun control opponent: "The right to bear arms is an important part of the constitution. Dems just want to take

away guns and our right to bear arms. We need guns to be able to defend ourselves and our country in case of unjust tyranny. The only thing that stops a bad guy with a gun is a good guy with a gun."
Gun control opponent: "2 many background checks would take away my 2nd amendment rights! What are we supposed to do, turn into Norway??!"
Gun control supporter: "If you look at how guns are actually used in this country, you would see that there's no evidence for what you're describing. THe more guns there are in an area, the more violence there is. Having guns doesn't make us any safer or more free."

Gun control opponent (message to be rephrased): "You communist! Having guns is an important part of my life as an American. It's one of the reasons I'm proud to be in this country. And I feel a million x safer when I have my gun with me"
Gun control supporter (message to respond to): "If you look at how guns are actually used in this country, you would see that there's no evidence for what you're describing. THe more guns there are in an area, the more violence there is. Having guns doesn't make us any safer or more free."
Gun control opponent (rephrased message with restatement): "I can see that you care about the safety of our country and decreasing violence and I respect that. Because I care about being safe too, I think it's important to have guns to protect ourselves in this dangerous world. Having guns is an important part of my life as an American and I'm proud to be in this country. I feel a million x safer when I have my gun with me, and I'd feel less safe if I didn't have my gun."
"""

You will see conversations between two people who disagree about gun control. Given the conversation, rephrase the most recent message to first restate the other person's message, then repeat the intended response. Also, if the language is very strong, try to soften the tone of the message. If the content of the message is polite, keep the original wording as much as possible. Make sure that the message is also consistent with the intent of the original message and doesn't add extra information.

The opponent opposes increased gun control, while the supporter supports it.

Conversation 3:
"""

Gun control opponent: "Bad people will always be able to get guns in this country. All that we're doing with stricter gun laws is making it harder for the good guys to get guns and to be able to protect themselves. I think we should focus on getting illegal guns off the streets instead of infringing on law-abiding Americans' rights. Plus, I like hunting with guns, and don't want that to be taken away from me."
Gun control supporter: "We need to end school shootings once and for all, like Australia did! Repeal the 2nd amendnment and buy back all the guns!"

Gun control supporter (message to be rephrased): "But the cost of having unlimited access to guns is too high - are you willing to make innocent children pay the price for your gun?"
Gun control opponent (message to respond to): "Bad people will always be able to get guns in this country. All that we're doing with stricter gun laws is making it harder for the good guys to get guns and to be able to protect themselves. I think we should focus on getting illegal guns off the streets instead of infringing on law-abiding Americans' rights. Plus, I like hunting with guns, and don't want that to be taken away from me."
Gun control supporter (rephrased message with restatement): "It seems like you think that keeping guns away from criminals is a good idea. I can see that you enjoy hunting too. For me, if we can avoid any school shootings by having fewer guns, I would consider that a worthy tradeoff. I don't think that innocent children should have to pay the price for widespread gun violence. What are your thoughts about that?"
"""

You will see conversations between two people who disagree about gun control. Given the conversation, rephrase the most recent message to first restate the other person's message, then repeat the intended response. Also, if the

language is very strong, try to soften the tone of the message. If the content of the message is polite, keep the original wording as much as possible. Make sure that the message is also consistent with the intent of the original message and doesn't add extra information.

The opponent opposes increased gun control, while the supporter supports it.

Conversation 4:

“““

<INSERT PREVIOUS TURNS IN THE CONVERSATION HERE>

<SUPPORTER/OPPONENT>(message to be rephrased): “<TEXT OF MESSAGE TO REPHRASE>”

<SUPPORTER/OPPONENT>(message to respond to): “<TEXT OF MESSAGE TO RESPOND TO>”

<SUPPORTER/OPPONENT>(rephrased message with restatement): “

C Sample Descriptive Characteristics and Imbalance Tests

C.1 Sample Demographics

Statistic	Mean	St. Dev.	Min	Median	Max
Age	48.273	14.035	18	48	88
Democrat	0.404	0.491	0	0	1
Republican	0.521	0.500	0	1	1
Independent	0.075	0.264	0	0	1
Female	0.570	0.495	0	1	1
Male	0.424	0.494	0	0	1
Non-binary	0.006	0.076	0	0	1
White	0.761	0.427	0	1	1
Black	0.130	0.337	0	0	1
Other Race/Ethnicity	0.109	0.311	0	0	1
Income	3.583	2.185	1	3	11
Education	4.107	1.908	1	4	9
Employment: Other	0.036	0.185	0	0	1
Retired	0.141	0.348	0	0	1
Student	0.022	0.148	0	0	1
Unemployed / looking	0.061	0.240	0	0	1
Unemployed/ not looking	0.058	0.233	0	0	1
Work full time	0.469	0.499	0	0	1
Work part time	0.112	0.316	0	0	1
Northeast	0.154	0.361	0	0	1
South	0.467	0.499	0	0	1
West	0.151	0.358	0	0	1
Suburban	0.414	0.493	0	0	1
Urban	0.272	0.445	0	0	1
Own Game Console	0.682	0.466	0	1	1

Table 4: Sample Demographic Summary Statistics

C.2 Randomization Balance Tests

In this section, we verify the randomization procedure by examining whether there are any substantial differences between the demographic profiles of the treatment and control groups. We estimate a linear probability model, using OLS to predict the binary assignment of respondents to a treatment or control conversation based on a number of their demographic characteristics.

Table 5 presents the results, which suggest a small difference in the employment variable, and another small difference in region in the final sample.

We take this as evidence that the randomization procedures were properly implemented. We also note that there is not a growing demographic imbalance between treatment and control among longer conversations, an important piece of evidence in favor of the balance assumption made in the placebo-controlled CACE analysis.

	Dependent variable: Binary Treatment Assignment				
	(0+)	(1+)	(2+)	(3+)	(4+)
Pre-Survey Support	0.01 (0.03)	0.01 (0.03)	0.03 (0.04)	0.04 (0.04)	0.03 (0.04)
Age	−0.001 (0.001)	−0.001 (0.001)	−0.0004 (0.001)	−0.001 (0.002)	0.0005 (0.002)
Party ID	0.001 (0.01)	−0.001 (0.01)	0.003 (0.01)	0.01 (0.01)	0.004 (0.01)
Male	0.01 (0.03)	0.01 (0.03)	0.01 (0.03)	0.01 (0.04)	0.02 (0.04)
Non-binary	0.07 (0.15)	0.07 (0.15)	−0.03 (0.17)	0.07 (0.21)	0.25 (0.23)
White	0.01 (0.03)	0.02 (0.03)	0.01 (0.04)	−0.01 (0.04)	−0.02 (0.05)
Income	−0.003 (0.01)	0.001 (0.01)	0.001 (0.01)	0.001 (0.01)	−0.01 (0.01)
Education	−0.001 (0.01)	−0.004 (0.01)	−0.003 (0.01)	−0.0001 (0.01)	0.0005 (0.01)
Employment: Other	0.11 (0.08)	0.11 (0.08)	0.14 (0.09)	0.11 (0.10)	0.10 (0.10)
Retired	0.08 (0.06)	0.04 (0.07)	−0.02 (0.07)	−0.01 (0.08)	−0.03 (0.09)
Student	0.16 (0.10)	0.22 (0.11)*	0.14 (0.12)	0.05 (0.14)	0.04 (0.16)
Unemployed/looking	0.15 (0.07)*	0.10 (0.07)	0.13 (0.08)	0.11 (0.10)	0.15 (0.10)
Unemployed/not looking	−0.04 (0.07)	−0.08 (0.07)	−0.13 (0.09)	−0.11 (0.10)	−0.10 (0.11)
Work full time	0.10 (0.05)*	0.10 (0.05)+	0.07 (0.06)	0.08 (0.06)	0.10 (0.07)
Work part time	0.08 (0.06)	0.07 (0.06)	0.04 (0.07)	0.05 (0.08)	0.05 (0.08)
Northeast	0.001 (0.04)	0.01 (0.05)	−0.02 (0.05)	−0.06 (0.06)	−0.03 (0.07)
South	0.03 (0.03)	0.04 (0.03)	0.05 (0.04)	0.06 (0.04)	0.09 (0.05)+
West	−0.001 (0.04)	−0.001 (0.05)	0.004 (0.05)	−0.02 (0.06)	−0.01 (0.07)
Suburban	0.004 (0.03)	0.01 (0.03)	0.02 (0.04)	0.01 (0.04)	0.005 (0.05)
Urban	−0.004 (0.04)	0.03 (0.04)	0.02 (0.04)	0.04 (0.05)	0.04 (0.05)
Own Game Console	−0.03 (0.03)	−0.03 (0.03)	−0.04 (0.04)	−0.02 (0.04)	−0.01 (0.05)
Constant	0.61 (0.09)**	0.60 (0.10)**	0.60 (0.11)**	0.56 (0.12)**	0.52 (0.13)**
Observations	1,445	1,263	938	782	653
R^2	0.01	0.02	0.02	0.02	0.03
Adjusted R^2	−0.004	−0.001	−0.003	−0.002	−0.002
Residual Std. Error	.47 (df = 1423)	.48 (df = 1241)	.47 (df = 916)	.48 (df = 760)	.47 (df = 631)
F Statistic	.76 (df=21; 1423)	.94 (df=21; 1241)	.88 (df=21; 916)	.91 (df=21; 760)	.95 (df=21; 631)
Note:					+p<0.1; *p<0.05; **p<0.01

Table 5: Randomization Balance Checks. OLS models predicting treatment assignment for each subgroup, based on a range of observable demographic characteristics. These models are run separately for each level of rephrasings received in the conversation.

C.3 Descriptive Statistics of the Conversations

In this section we provide some descriptive statistics about the conversations and rephrasing interventions.

- Total Number of Messages Sent: 25,612
- Average Messages Per Conversation: 12
  - 1 rephrasings: 6
  - 2 rephrasings: 13
  - 3 rephrasings: 18
  - 4 rephrasings: 23
- Total Number of Rephrasing Interventions: 2,742
- Total Number of Rephrasings Accepted: 1,798 (66%)
  - Polite Rephrasings Accepted: 724 (40%)
  - Restate Rephrasings Accepted: 530 (30%)
  - Validate Rephrasings Accepted: 544 (30%)
- Number of Rephrasing Suggestions Edited by the User: 9

D Attrition and Dosage

Figures 8 and 9 present two visualizations that illustrate when respondents dropped out of the study and which participants received what dosage levels of the interventions (GPT-3 rephrasings).

E Analysis of Conversation Content

As noted in the main text and in the methods sections of these Supplementary Materials, we explored the content of the messages in the conversations in various ways. These analyses has several objectives, such as ensuring that the rephrasings from GPT-3 differed from users' statements in the ways we intended, determining if the conversations themselves focused on the assigned subject of gun regulation, and considering how the suggestions from GPT-3 shifted the nature of subsequent statements in the chat rooms.

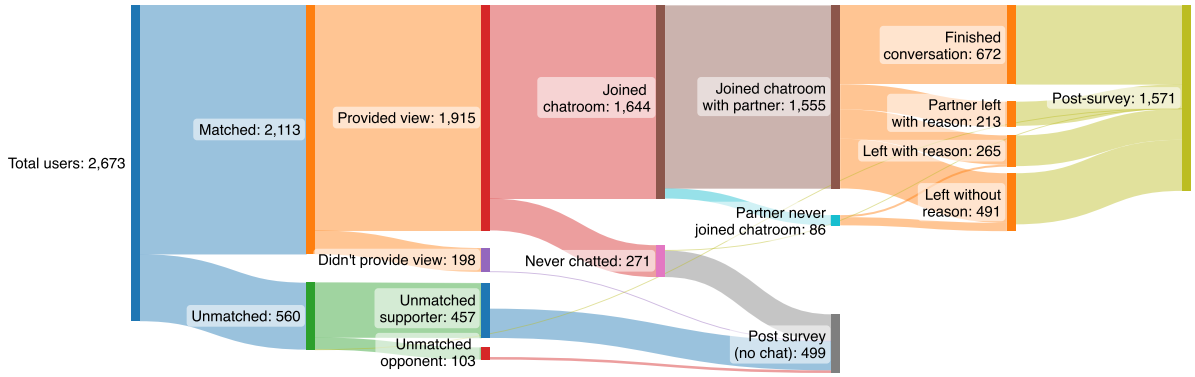

Figure 8: A sankey diagram showing attrition and respondent outcomes on the chat platform. All data for this diagram comes from within the chat platform, and not from the surveys. 3 individuals who took the post-survey were not matched to a user ID from the chat platform.

E.1 Analyzing Tone of Rewritten Messages

Table 3 presented earlier in the Methods section indicates that the rephrasings created by GPT-3 and chosen by participants differed significantly in tone from the users' original statements in ways intended. This analysis relied on coding by the Stanford politeness package, JigSaw/Google's Perspective API, and Google's Cloud NLP sentiment analysis API. For each API, we analyzed each message that had an accepted rephrasing with the default parameters and no additional text processing.

Significance was determined using Scipy's standard paired t-test. While none of our conversations were very toxic, according to almost every metric, we see that our rewrites either have no effect or move the tone of the message towards more polite and civil language. It is difficult for these automated APIs to measure the effects of subtle lexical changes, and yet we see that they have significant effects on overall conversational outcomes; finding better ways to quantify which features, exactly, of our rewrites were most impactful is an important step for future research.

E.2 Analyzing the Impact of Rewrites on Overall Conversational Tone

Figure 5 in the main text demonstrates how the rephrasings from GPT-3 had downstream effects on the later tone and characteristics of the chats between participants. We used the following procedure to generate it:

- For each treated conversation, we split the conversation's messages into two groups: those that happened before (but not including) the first accepted intervention, and those that happened after. We denote these the "pre-" set and the "post-" set of messages. These

conversations were usually split on the 4th message; some were split later because of the technical details of how turns were counted and how we decided when to offer a rephrasing.

- For each untreated conversation, we likewise split the conversation into pre- and post-halves. Here, we simply split on the 4th message, as this was the median splitpoint of treated conversations.
- For each set of pre- and post- messages, we analyzed each message using the Perspective API, and calculated the mean and variance of each set.
- Significance was determined using scipy's standard two-tailed t-test.

We did not include any GPT-3 generated rephrasings in this analysis, allowing us to isolate just the effect of rewrites on *other* messages. The analyses in this figure suggest that the interventions from GPT-3 shifted the tone of the conversations in ways aligned with our experimental design and prompting of GPT-3. For example, consider the "Toxicity" measure: our results suggest that messages before turn 4 in both treated and untreated conversations have similar levels of toxicity, but that messages after interventions are significantly less toxic. This suggests that our rephrasings create a virtuous cycle of improved tone.

E.3 Analyzing the Distribution of Conversational Topics and Rewrites

Figure 2 in the main text (panels A and B) considers the content of the messages sent in the chatrooms. To create this visualization and perform this analysis, we converted all 10,695 messages over 4 words long into a 768-dimensional feature vector using the `sentence-tranformers` library and the `sentence-t5-xxl` model. This model is optimized for capturing semantic similarity between sentences. In line with standard practice, these large-dimensional vectors were then reduced to 50 dimensions with PCA and embedded into the 2D space using the UMAP library (with n_neighbors=5, min_dist=0.001, and a cosine similarity metric). The resulting two-dimensional points were then clustered using `scikit-learn`'s AgglomerativeClustering algorithm, which is an unsupervised clustering algorithm. We used default parameters, except that we asked it to create 25 clusters. To summarize the content of each cluster without introducing any human bias, we used GPT-3 (`text-davinci-003`) to automatically generate a synthetic label for the cluster based on a simple prompt. These labels were included as-generated, without any cherry picking or additional text processing. This creates the cluster names shown in Figure 2. Our own manual checking of this clustering supported the labels created by GPT-3.

Panel A of Figure 2 shows that the topics participants discussed were on the subject we intended (gun regulation), although the specific aspect of gun regulation varied across the messages. Panel B shows that the rephrased messages from GPT-3 were spread across these clusters. Had the rephrasings been located primarily in a single cluster or been focused in only one part

of this figure, we might have concerns about the rephrasings changing the content and not just the politeness, validation, and restating we intended.

F Factor Analysis of Index Items

The Methods section describes how we combined various items into index dependent variables. In addition to calculating Cronbach's α, we also conducted exploratory factor analyses for each of these measures. The EFA results lend additional support for the combination of the individual items into the indices used in the paper. Plots for the very simple structure fit of these sets of questions support a single index for each concept, as shown in figures 10, 11, 12, and 13. In these figures, a single factor performs generally as well as a two or more factor solution, confirming the results of the Cronbach's α estimates shown in the Methods section.

Additional details about the specific factor loadings from these analyses can be found in Table 6 for the conversation items, Table 7 for the divisiveness items, Table 8 for the gun policy items on the pre-chat survey, and Table 9 for the gun policy items on the post-chat survey.

Item	Loadings
Grade	0.7469
Stressful	0.3854
Felt Understood	0.8512
Treated w/ Respect	0.4781
Disrespectful	0.5476
Change views	0.3404
communicate	0.7067

Table 6: Factor loadings for a one-factor exploratory factor analysis of the conversation items. All items are scored so that higher values indicate better conversations.

Item	Loadings
Feeling thermometer	0.3273
Point of view	0.4856
Perspective	0.7237
Good reasons	0.7465
Respect	0.7893

Table 7: Factor loadings for a one-factor exploratory factor analysis of the divisiveness items. All items are scored so that higher values indicate more divisiveness.

Item	Loadings
Mental Illness Purchases	0.4906
Ban Assault-style	0.8379
Ban High-capacity	0.8211
Carry Concealed	0.5593
Arm Teachers	0.4273
Enhanced Background checks	0.6228
Red Flag Laws	0.7275

Table 8: Factor loadings for a one-factor exploratory factor analysis of the gun policy items on the pre-chat survey. All items are scored so that higher values indicate more support for gun restrictions.

Item	Loadings
Mental Illness Purchases	0.537
Ban Assault-style	0.8429
Ban High-capacity	0.8312
Carry Concealed	0.5688
Arm Teachers	0.3856
Enhanced Background checks	0.6559
Red Flag Laws	0.7304

Table 9: Factor loadings for a one-factor exploratory factor analysis of the gun policy items on the post-chat survey. All items are scored so that higher values indicate more support for gun restrictions.

G Tests of the Placebo Control CACE Assumption

One of the central features of the analyses presented in the main text is our presentation of results by treatment dosage/exposure, measured by conversation length. As we note in the paper, causal identification for subgroups that had longer conversations (meaning more treatment exposure) hinges on the assumption that respondents' chat length is unrelated to the treatment condition. Here we test that assumption by exploring what characteristics, if any, correlate with levels of treatment dosage/exposure. In this analysis, we use a linear regression model to predict the number of rephrasings offered in a conversation (or the number of rephrasings that would have been offered in a control conversation had it been treated) by treatment group assignment, pre-treatment support for gun regulation, age, party ID, gender, race, income, education, employment, region, neighborhood, and game console ownership (a proxy for technology comfort).

As Table 10 indicates, we do find some small but statistically significant correlations between conversation length and respondents' pre-chat survey attitudes, party ID, race, income, and education. Notably, however, treatment assignment is *not* a significant predictor of length of

conversation. We measure length of conversation in two ways; in models 1 and 2 and it is the number of rephrasings shown (or the number that would have been shown in a control conversation had it been treated), and in models 3 and 4 it is the total number of messages sent by either partner in the conversation.

H Additional Treatment Effect Results

In the main text, we rely on a series of indices for our analyses instead of examining the constituent items individually. We do this for the conversation quality measures, the divisiveness questions, and attitudes about gun policy. This approach recognizes the various benefits provided by relying on multi-item scales (*64, 65*), and is supported by the psychometric properties of these items as described in Section F. We also rely on these indices for ease and simplicity of presentation in the main text of the paper.

In this part of the appendix, we present the results separately for each item that makes up each index. For each index, we first present a figure that graphs the means and 95% confidence intervals for each variable by treatment condition, which covers the ITT and placebo-control CACE estimates. We then present a numerical table of the means and standard deviations used to create those figures. Third, we present the coefficient and p-value from the two-stage least squares CACE model for each combination of variable, treatment condition, and definition of compliance. Finally, for just the conversational quality and divisiveness indices, we present an alternative version of the two-stage least squares CACE analysis that uses the number of rephrasings *accepted into the conversation* as the metric for treatment exposure, rather than just the number of times a rephrasing was offered.

For just the Divisiveness Index items, we also present linear regression models which provide a statistical test of the difference in means (placebo-control CACE) displayed in the figures.

	Dependent variable:			
	Rephrasing Interventions Shown		Total Messages in Convo	
	(1)	(2)	(3)	(4)
Treat: GPT-3 partner	−0.01 (0.10)	−0.04 (0.11)	0.76 (0.58)	0.63 (0.61)
Treat: GPT-3 intervention	0.01 (0.10)	−0.03 (0.11)	0.68 (0.58)	0.50 (0.61)
Gun Control Supporter		0.18 (0.10)$^{+}$		1.01 (0.57)$^{+}$
Age		0.01 (0.004)		0.01 (0.02)
Party ID		0.04 (0.02)$^{+}$		0.20 (0.12)
Male		0.02 (0.09)		0.43 (0.54)
Non-binary		0.47 (0.51)		1.15 (2.91)
Black		−0.35 (0.15)*		−1.65 (0.84)*
Other Race		0.20 (0.14)		1.04 (0.81)
Employment: Other		0.03 (0.26)		0.88 (1.52)
Employment: Retired		−0.17 (0.21)		−0.79 (1.20)
Employment: Student		−0.26 (0.34)		−1.07 (1.95)
Unemployed - Looking		−0.28 (0.23)		−2.16 (1.32)
Unemployed - Not Looking		−0.54 (0.23)*		−3.15 (1.33)*
Employed - Full time		−0.21 (0.16)		−1.49 (0.93)
Employed - Part time		−0.22 (0.20)		−1.89 (1.13)$^{+}$
Income		0.04 (0.02)		0.16 (0.14)
Education		0.05 (0.03)$^{+}$		0.38 (0.16)*
Northeast		0.002 (0.15)		0.29 (0.85)
South		0.11 (0.11)		0.66 (0.64)
West		0.005 (0.15)		0.05 (0.86)
Suburban		−0.004 (0.11)		−0.07 (0.61)
Urban		−0.07 (0.12)		−0.59 (0.71)
Owns Game Console		−0.09 (0.11)		−0.32 (0.62)
Constant	2.56 (0.07)**	2.01 (0.32)**	15.02 (0.41)**	12.66 (1.82)**
Observations	1,574	1,445	1,574	1,445
R^2	0.0000	0.03	0.001	0.03
Adjusted R^2	−0.001	0.01	0.0000	0.01
Residual Std. Error	1.64 (df=1571)	1.64 (df=1420)	9.44 (df=1571)	9.45 (df=1420)
F Statistic	0.01 (df=2; 1571)	1.90** (df=24; 1420)	1.03 (df=2; 1571)	1.83** (df=24; 1420)
Note:				$^{+}$p<0.1; *p<0.05; **p<0.01

Table 10: OLS Models exploring Balance Assumption

H.1 Conversation Quality Index

		Control		GPT-3 Self				GPT-3 Partner			
		Mean	n	Mean	n	Diff	p	Mean	n	Diff	p
1	(ITT) 0+	0.71	530	0.71	521	0.00	0.81	0.73	523	0.02	0.04
2	1+	0.72	478	0.73	448	0.01	0.45	0.76	447	0.03	0.00
3	2+	0.76	330	0.77	343	0.01	0.63	0.78	345	0.02	0.08
4	3+	0.77	283	0.78	280	0.01	0.38	0.80	277	0.03	0.03
5	4+	0.77	233	0.78	233	0.01	0.35	0.80	232	0.04	0.01

Table 11: Conversation Index T-Tests: Columns indicate the mean and sample size for the conversation quality index by treatment condition. The mean difference and p-value for the treatment conditions are relative to the control condition, and are derived from a standard independent-samples t-test. These are the significance levels reported in the main text Figure 3.

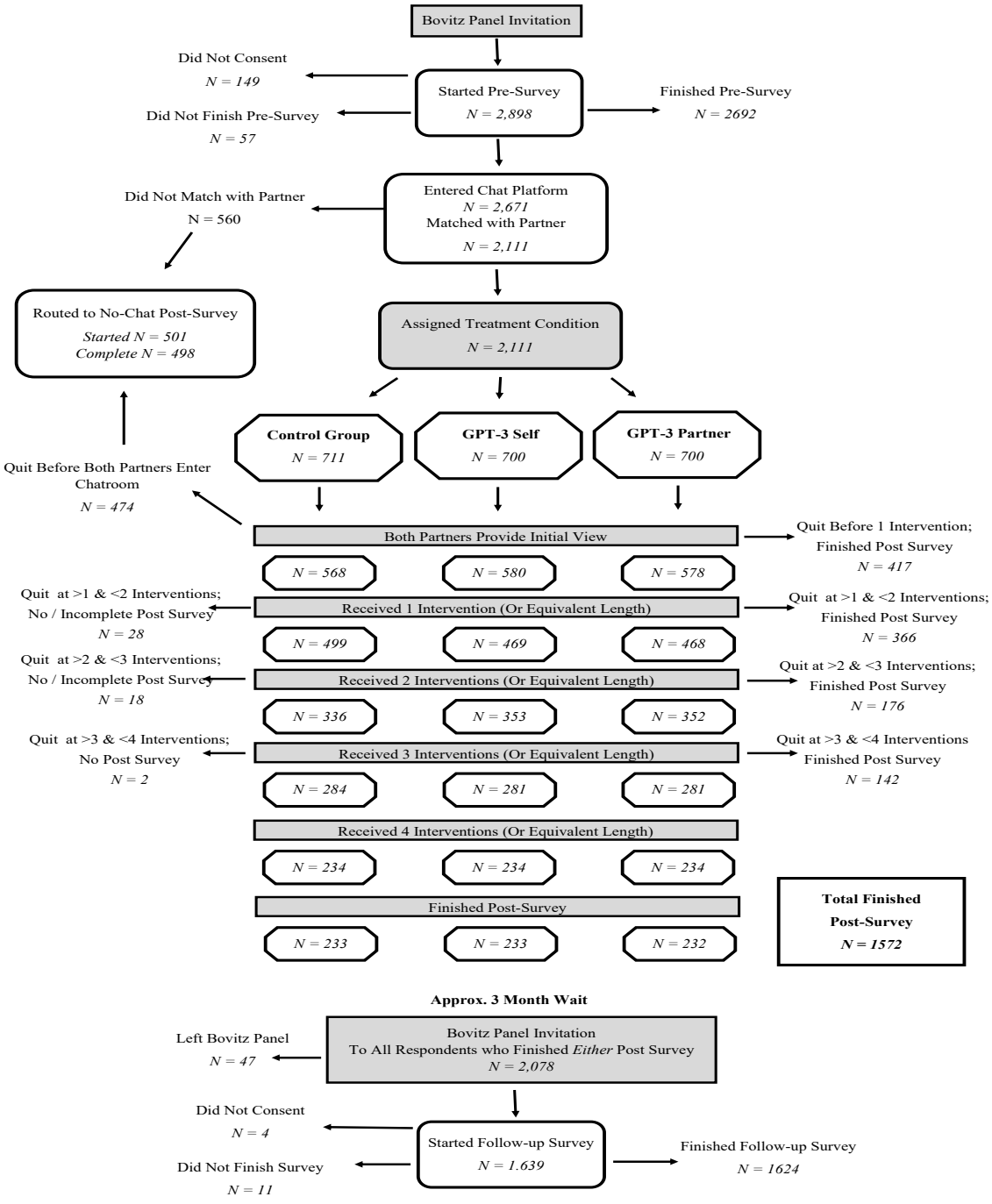

Figure 9: A diagram showing attrition, dosage, and treatment assignment over the full life-cycle of the experiment. 2 people started, but did not fully complete, the post-survey. We include these individuals in the analysis, leading to an effective total sample size of 1,574.

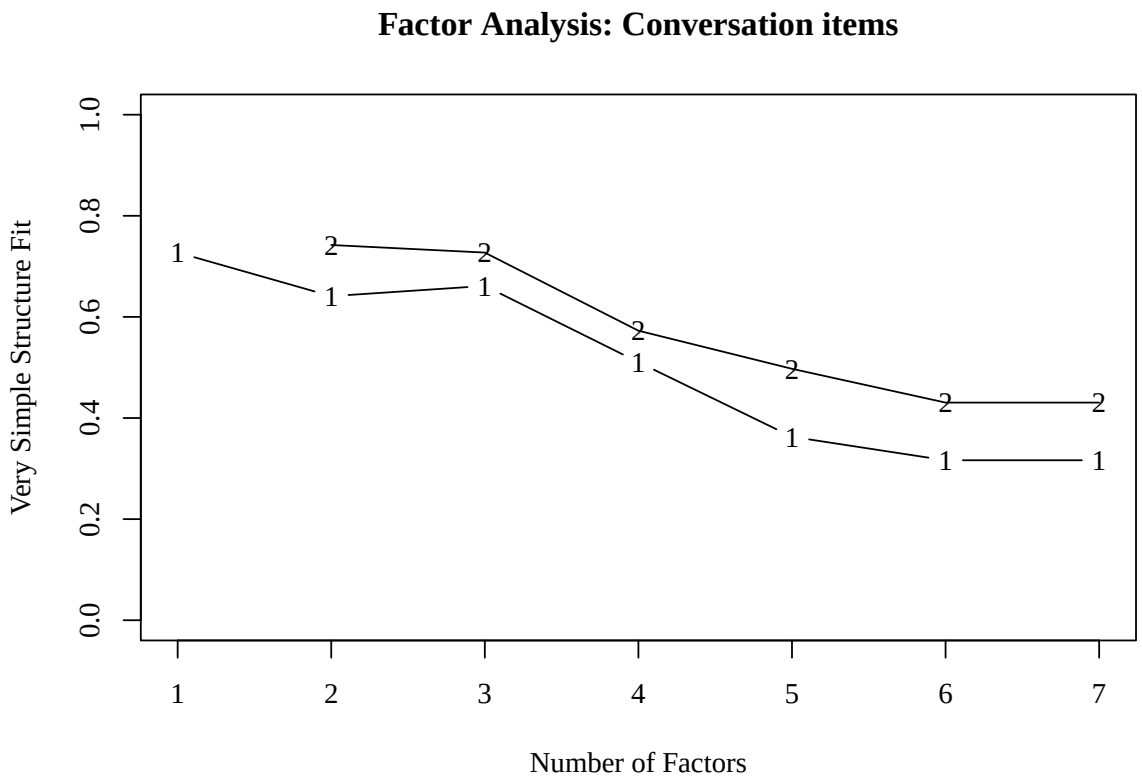

Figure 10: Based on factor analysis of conversation quality items

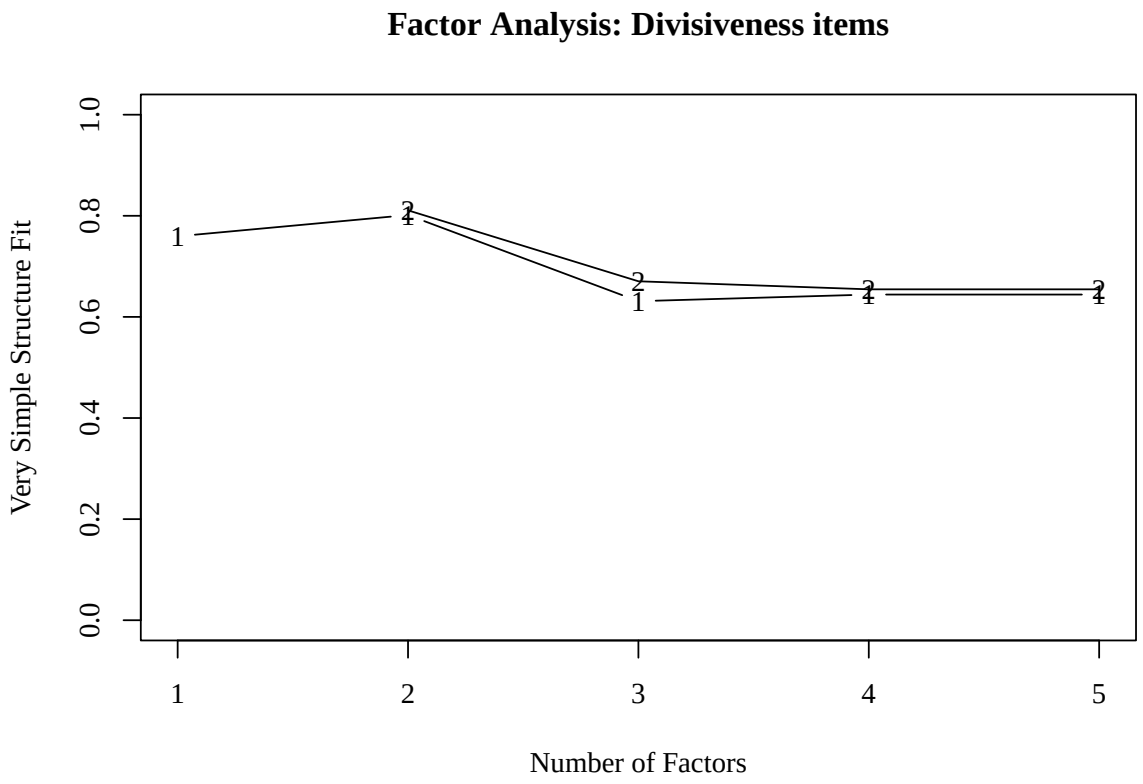

Figure 11: Based on factor analysis of divisiveness items

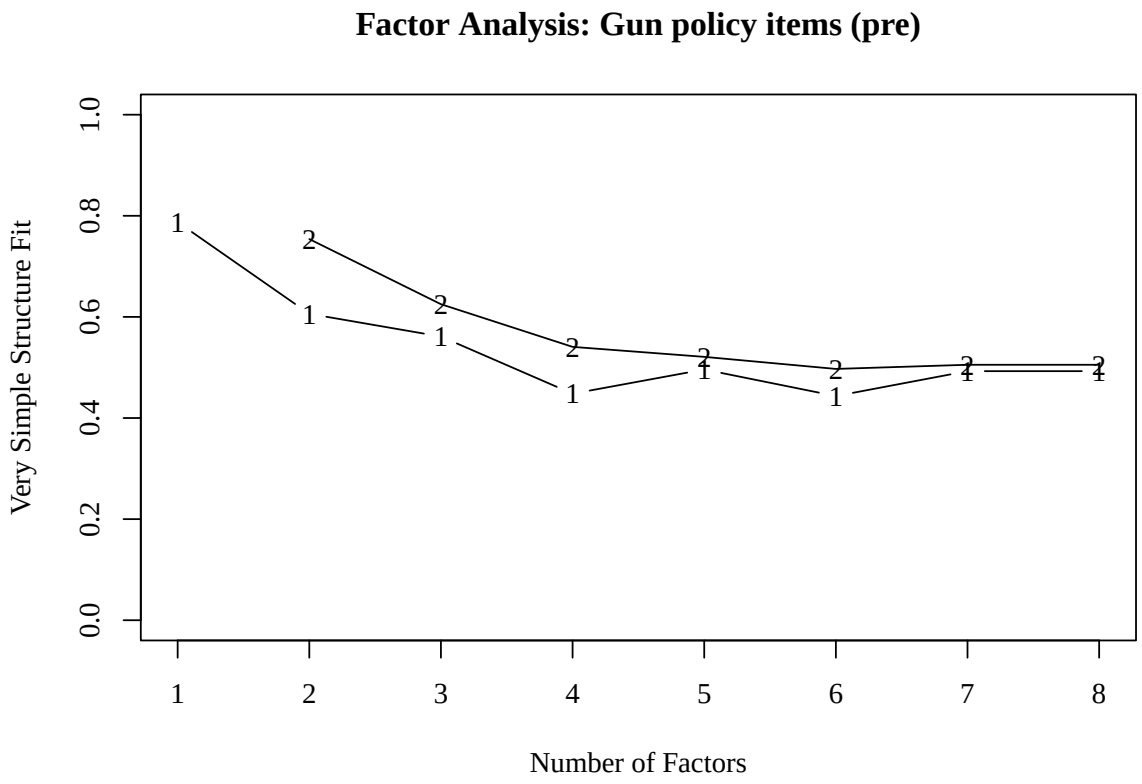

Figure 12: Based on factor analysis of pre-treatment gun policy questions

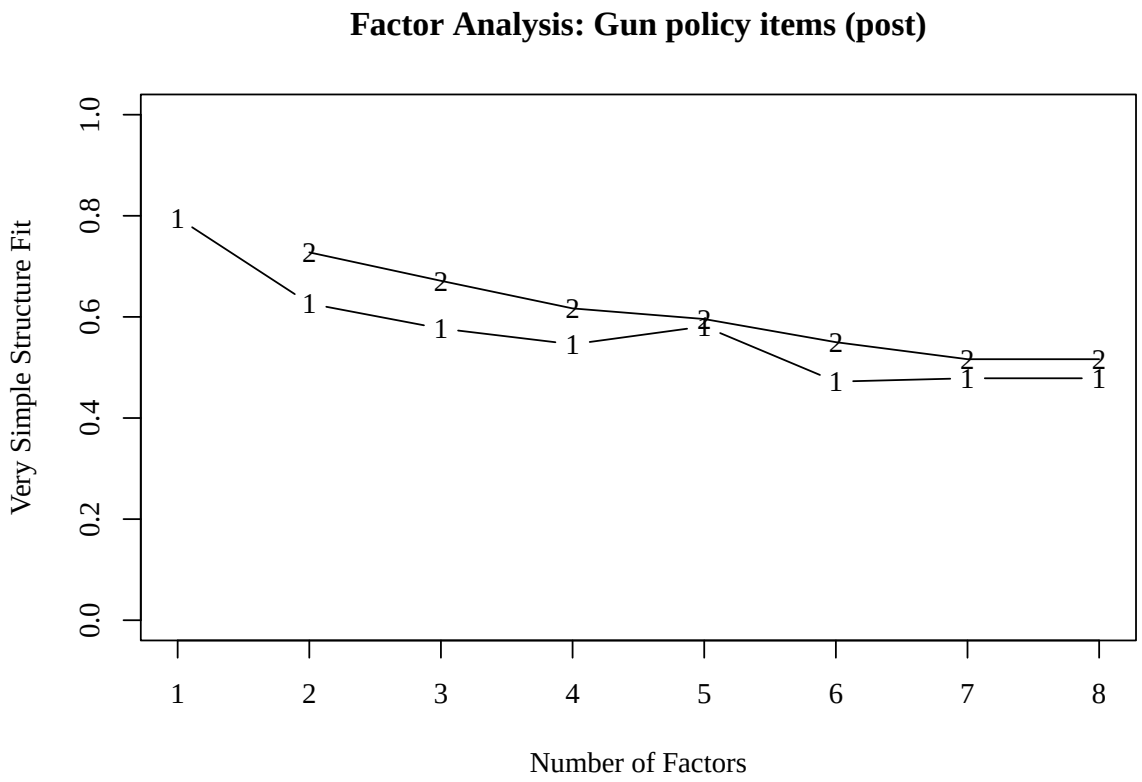

Figure 13: Based on factor analysis of post-treatment gun policy questions

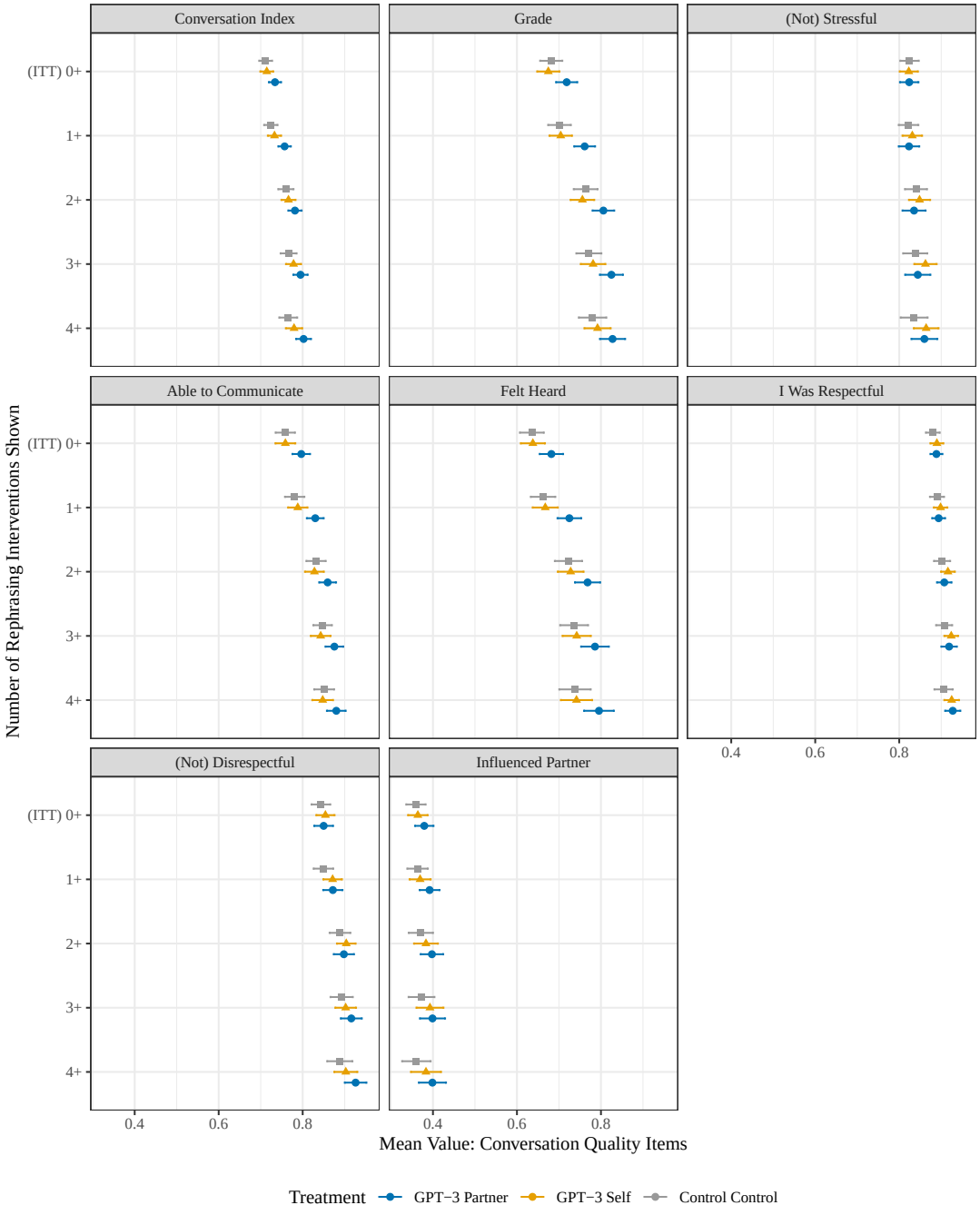

Figure 14: Placebo-control CACE Effects: Mean and 95% confidence interval for all metrics that make up the conversation index, by treatment condition and conversation length.

	DV	Treatment	0+ mean	0+ sd	1+ mean	1+ sd	2+ mean	2+ sd	3+ mean	3+ sd	4+ mean	4+ sd
1	Conversation Index	Control	0.71	0.18	0.72	0.18	0.76	0.17	0.77	0.16	0.77	0.17
2	Conversation Index	Partner	0.73	0.17	0.76	0.16	0.78	0.15	0.79	0.15	0.80	0.14
3	Conversation Index	Self	0.71	0.18	0.73	0.17	0.77	0.16	0.78	0.15	0.78	0.15
4	Grade	Control	0.68	0.31	0.70	0.30	0.76	0.26	0.77	0.26	0.78	0.25
5	Grade	Partner	0.72	0.30	0.76	0.27	0.81	0.24	0.82	0.23	0.83	0.23
6	Grade	Self	0.67	0.31	0.70	0.29	0.76	0.27	0.78	0.25	0.79	0.24
7	(Not) Stressful	Control	0.82	0.26	0.82	0.26	0.84	0.24	0.84	0.25	0.84	0.25
8	(Not) Stressful	Partner	0.82	0.26	0.82	0.26	0.84	0.26	0.84	0.25	0.86	0.24
9	(Not) Stressful	Self	0.82	0.25	0.83	0.25	0.85	0.24	0.86	0.23	0.86	0.23
10	Able to Communicate	Control	0.76	0.27	0.78	0.26	0.83	0.21	0.85	0.19	0.85	0.18
11	Able to Communicate	Partner	0.80	0.24	0.83	0.22	0.86	0.19	0.88	0.18	0.88	0.17
12	Able to Communicate	Self	0.76	0.28	0.79	0.25	0.83	0.21	0.84	0.20	0.85	0.19
13	Felt Heard	Control	0.64	0.34	0.66	0.33	0.72	0.30	0.74	0.28	0.74	0.29
14	Felt Heard	Partner	0.68	0.33	0.72	0.30	0.77	0.28	0.79	0.28	0.80	0.27
15	Felt Heard	Self	0.64	0.33	0.67	0.32	0.73	0.29	0.74	0.29	0.74	0.29
16	I Was Respectful	Control	0.88	0.19	0.89	0.19	0.90	0.18	0.91	0.16	0.91	0.17
17	I Was Respectful	Partner	0.89	0.17	0.89	0.17	0.91	0.16	0.92	0.16	0.93	0.14
18	I Was Respectful	Self	0.89	0.18	0.90	0.17	0.92	0.16	0.92	0.14	0.92	0.14
19	(Not) Disrespectful	Control	0.84	0.26	0.85	0.26	0.89	0.23	0.89	0.23	0.89	0.23
20	(Not) Disrespectful	Partner	0.85	0.26	0.87	0.25	0.90	0.23	0.92	0.21	0.93	0.20
21	(Not) Disrespectful	Self	0.85	0.25	0.87	0.24	0.90	0.21	0.90	0.21	0.90	0.21
22	Influenced Partner	Control	0.36	0.27	0.36	0.27	0.37	0.27	0.37	0.26	0.36	0.26
23	Influenced Partner	Partner	0.38	0.26	0.39	0.26	0.40	0.25	0.40	0.25	0.40	0.25
24	Influenced Partner	Self	0.36	0.27	0.37	0.27	0.38	0.27	0.39	0.27	0.38	0.28

Table 12: Conversation Index Item Means: Columns indicate the mean and standard deviation for the conversation quality index, followed by each constituent item, by subsets of the minimum number of rephrasings that was shown to the treated respondent (or that would have been shown in a control condition if it had been treated.)

			1+ Interventions		Partial Compliance		4+ Interventions	
	DV	Treatment	AME	p	AME	p	AME	p
1	Conversation Index	GPT-3 Partner	0.03	0.05	0.04	0.05	0.05	0.05
2	(Not) Stressful	GPT-3 Partner	0.00	0.98	0.00	0.98	0.00	0.98
3	Able to Communicate	GPT-3 Partner	0.04	0.03	0.06	0.03	0.08	0.03
4	Felt Heard	GPT-3 Partner	0.06	0.03	0.08	0.03	0.11	0.03
5	I Was Respectful	GPT-3 Partner	0.01	0.40	0.01	0.40	0.02	0.40
6	(Not) Disrespectful	GPT-3 Partner	0.01	0.64	0.01	0.64	0.02	0.64
7	Influenced Partner	GPT-3 Partner	0.02	0.22	0.03	0.22	0.04	0.23
8	Conversation Index	GPT-3 Self	0.00	0.95	0.00	0.95	0.00	0.95
9	(Not) Stressful	GPT-3 Self	-0.00	0.84	-0.00	0.84	-0.01	0.84
10	Able to Communicate	GPT-3 Self	-0.00	0.92	-0.00	0.92	-0.00	0.92
11	Felt Heard	GPT-3 Self	-0.00	1.00	-0.00	1.00	-0.00	1.00
12	I Was Respectful	GPT-3 Self	0.01	0.53	0.01	0.53	0.02	0.53
13	(Not) Disrespectful	GPT-3 Self	0.01	0.62	0.01	0.62	0.02	0.62
14	Influenced Partner	GPT-3 Self	0.00	0.82	0.01	0.82	0.01	0.82

Table 13: Conversation Quality TSLS CACE Analysis: Cell values are average marginal effects (coefficients) and p-values calculated from a two-stage least squares CACE model, with standard errors clustered by conversation. Instruments are treatment assignment and a binary indicator for whether the partner sent a single message. Controls are pre-chat position on gun control, party ID, employment, education, and race.

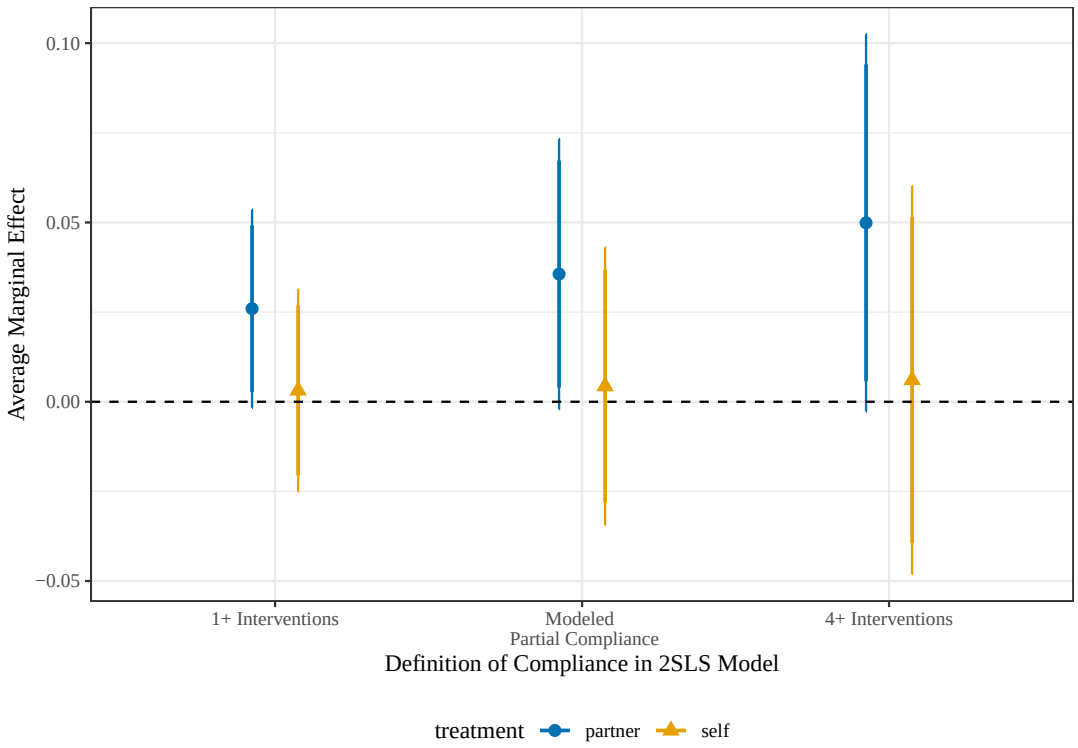

Figure 15: Alternative Conversation Quality TSLS CACE results, with no controls. This basic model includes only the treatment assignment as an instrument for treatment dosage.

H.2 Divisiveness Index

		Control		GPT-3 Self				GPT-3 Partner			
		Mean	n	Mean	n	Diff	p	Mean	n	Diff	p
1	(ITT) 0+	0.32	530.00	0.31	521.00	-0.01	0.53	0.30	523.00	-0.02	0.12
2	1+	0.31	478.00	0.30	448.00	-0.01	0.38	0.28	447.00	-0.03	0.02
3	2+	0.31	330.00	0.28	343.00	-0.03	0.03	0.27	345.00	-0.04	0.00
4	3+	0.32	283.00	0.27	280.00	-0.04	0.01	0.26	277.00	-0.05	0.00
5	4+	0.32	233.00	0.28	233.00	-0.04	0.01	0.25	232.00	-0.07	0.00

Table 14: Divisiveness Index T-Tests: Columns indicate the mean and sample size for the divisiveness index by treatment condition. The mean difference and p-value for the treatment conditions are relative to the control condition, and are derived from a standard independent-samples t-test. These are the significance levels reported in the main text Figure 4.

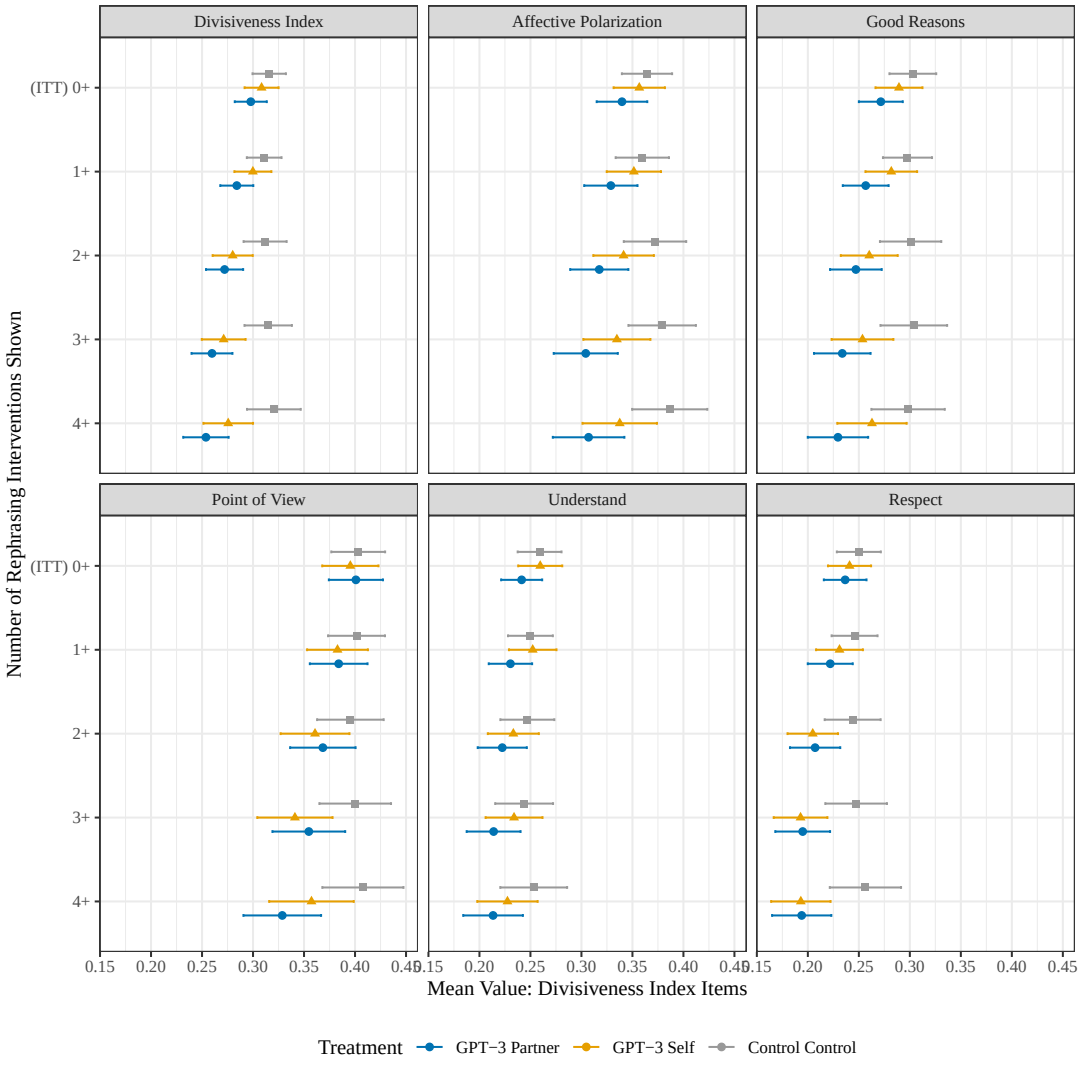

Figure 16: Mean and 95% confidence interval for all metrics that make up the divisiveness index, by treatment condition and conversation length.

	DV	Treatment	0+ mean	0+ sd	1+ mean	1+ sd	2+ mean	2+ sd	3+ mean	3+ sd	4+ mean	4+ sd
1	Divisiveness Index	Control	0.32	0.19	0.31	0.19	0.31	0.19	0.31	0.20	0.32	0.20
2	Divisiveness Index	Partner	0.30	0.18	0.28	0.17	0.27	0.17	0.26	0.17	0.25	0.17
3	Divisiveness Index	Self	0.31	0.19	0.30	0.20	0.28	0.19	0.27	0.18	0.28	0.19
4	Affective Polarization	Control	0.36	0.29	0.36	0.29	0.37	0.28	0.38	0.28	0.39	0.29
5	Affective Polarization	Partner	0.34	0.29	0.33	0.28	0.32	0.27	0.30	0.27	0.31	0.27
6	Affective Polarization	Self	0.36	0.29	0.35	0.29	0.34	0.28	0.33	0.28	0.34	0.28
7	Good Reasons	Control	0.30	0.27	0.30	0.27	0.30	0.28	0.30	0.28	0.30	0.28
8	Good Reasons	Partner	0.27	0.25	0.26	0.24	0.25	0.24	0.23	0.23	0.23	0.23
9	Good Reasons	Self	0.29	0.27	0.28	0.27	0.26	0.26	0.25	0.26	0.26	0.26
10	Point of View	Control	0.40	0.31	0.40	0.31	0.40	0.30	0.40	0.30	0.41	0.31
11	Point of View	Partner	0.40	0.31	0.38	0.30	0.37	0.30	0.35	0.30	0.33	0.29
12	Point of View	Self	0.40	0.32	0.38	0.32	0.36	0.32	0.34	0.31	0.36	0.32
13	Understand	Control	0.26	0.25	0.25	0.24	0.25	0.24	0.24	0.24	0.25	0.25
14	Understand	Partner	0.24	0.23	0.23	0.23	0.22	0.23	0.21	0.22	0.21	0.23
15	Understand	Self	0.26	0.25	0.25	0.25	0.23	0.24	0.23	0.24	0.23	0.23
16	Respect	Control	0.25	0.25	0.25	0.25	0.24	0.25	0.25	0.26	0.26	0.27
17	Respect	Partner	0.24	0.24	0.22	0.24	0.21	0.23	0.19	0.23	0.19	0.22
18	Respect	Self	0.24	0.25	0.23	0.25	0.20	0.23	0.19	0.22	0.19	0.23

Table 15: Divisiveness Index Means: Columns indicate the mean and standard deviation for the divisiveness index, followed by each constituent item, by subsets of the minimum number of rephrasings that was shown to the treated respondent (or that would have been shown in a control condition if it had been treated.)

For the Divisiveness Index and each of the constituent variables, we also provide a simple linear regression model that predicts each outcome with the treatment assignment as the sole independent variable. This functions as a statistical test of the mean differences visualized in Figure 4 in the main paper.

	Dependent variable:				
	Divisiveness Index				
	0+	1+	2+	3+	4+
Partner	-0.02	-0.03^{*}	-0.04^{**}	-0.06^{**}	-0.07^{**}
	(0.01)	(0.01)	(0.01)	(0.02)	(0.02)
Self	-0.01	-0.01	-0.03^{*}	-0.04^{**}	-0.04^{*}
	(0.01)	(0.01)	(0.01)	(0.02)	(0.02)
Constant	0.32^{**}	0.31^{**}	0.31^{**}	0.31^{**}	0.32^{**}
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Observations	1,571	1,370	1,016	839	697
R^2	0.002	0.004	0.01	0.02	0.02
Adjusted R^2	0.0003	0.002	0.01	0.01	0.02
Note:					$^{+}$p<0.1; *p<0.05; **p<0.01

Table 16: Divisiveness Index OLS Models

	Dependent variable:				
	Affective Polarization				
	0+	1+	2+	3+	4+
Partner	-0.02	-0.03	-0.05^{*}	-0.07^{**}	-0.08^{**}
	(0.02)	(0.02)	(0.02)	(0.02)	(0.03)
Self	-0.01	-0.01	-0.03	-0.04^{+}	-0.05^{+}
	(0.02)	(0.02)	(0.02)	(0.02)	(0.03)
Constant	0.36^{**}	0.36^{**}	0.37^{**}	0.38^{**}	0.39^{**}
	(0.01)	(0.01)	(0.02)	(0.02)	(0.02)
Observations	1,571	1,370	1,016	839	697
R^2	0.001	0.002	0.01	0.01	0.01
Adjusted R^2	-0.0000	0.001	0.004	0.01	0.01
Note:					$^{+}$p<0.1; *p<0.05; **p<0.01

Table 17: Affective Polarization OLS Models

	Dependent variable:				
	Good Reasons				
	0+	1+	2+	3+	4+
Partner	-0.03^{+}	-0.04^{*}	-0.05^{**}	-0.07^{**}	-0.07^{**}
	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
Self	-0.01	-0.02	-0.04^{*}	-0.05^{*}	-0.04
	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
Constant	0.30^{**}	0.30^{**}	0.30^{**}	0.30^{**}	0.30^{**}
	(0.01)	(0.01)	(0.01)	(0.02)	(0.02)
Observations	1,572	1,371	1,017	840	698
R^2	0.002	0.004	0.01	0.01	0.01
Adjusted R^2	0.001	0.003	0.01	0.01	0.01
Note:					$^{+}$p<0.1; *p<0.05; **p<0.01

Table 18: Good Reasons OLS Models

	Dependent variable:				
	Understand				
	0+	1+	2+	3+	4+
Partner	-0.02	-0.02	-0.02	-0.03	-0.04^{+}
	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
Self	0.001	0.002	-0.01	-0.01	-0.03
	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
Constant	0.26^{**}	0.25^{**}	0.25^{**}	0.24^{**}	0.25^{**}
	(0.01)	(0.01)	(0.01)	(0.01)	(0.02)
Observations	1,572	1,371	1,017	840	698
R^2	0.001	0.002	0.002	0.003	0.005
Adjusted R^2	-0.0001	0.0002	-0.0002	0.0004	0.002
Note:					$^{+}$p<0.1; *p<0.05; **p<0.01

Table 19: Understand OLS Models

	Dependent variable:				
	Point of View				
	0+	1+	2+	3+	4+
Partner	-0.002	-0.02	-0.03	-0.05^{+}	-0.08^{**}
	(0.02)	(0.02)	(0.02)	(0.03)	(0.03)
Self	-0.01	-0.02	-0.03	-0.06^{*}	-0.05^{+}
	(0.02)	(0.02)	(0.02)	(0.03)	(0.03)
Constant	0.40^{**}	0.40^{**}	0.40^{**}	0.40^{**}	0.41^{**}
	(0.01)	(0.01)	(0.02)	(0.02)	(0.02)
Observations	1,572	1,371	1,017	840	698
R^2	0.0001	0.001	0.002	0.01	0.01
Adjusted R^2	-0.001	-0.001	0.0003	0.005	0.01
Note:					$^{+}$p<0.1; *p<0.05; **p<0.01

Table 20: Point of View OLS Models

	Dependent variable:				
	Respect				
	0+	1+	2+	3+	4+
Partner	-0.01	-0.02	-0.04^{*}	-0.05^{**}	-0.06^{**}
	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
Self	-0.01	-0.01	-0.04^{*}	-0.05^{**}	-0.06^{**}
	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
Constant	0.25^{**}	0.25^{**}	0.24^{**}	0.25^{**}	0.26^{**}
	(0.01)	(0.01)	(0.01)	(0.01)	(0.02)
Observations	1,572	1,371	1,017	840	698
R^2	0.001	0.002	0.01	0.01	0.01
Adjusted R^2	-0.001	0.0002	0.004	0.01	0.01
Note:					$^{+}$p<0.1; *p<0.05; **p<0.01

Table 21: Respect OLS Models

			1+ Interventions		Partial Compliance		4+ Interventions	
	DV	Treatment	AME	p	AME	p	AME	p
1	Divisiveness Index	GPT-3 Partner	-0.02	0.08	-0.03	0.07	-0.05	0.07
2	Affective Polarization	GPT-3 Partner	-0.03	0.14	-0.04	0.14	-0.06	0.14
3	Good Reasons	GPT-3 Partner	-0.04	0.03	-0.06	0.03	-0.08	0.03
4	Point of View	GPT-3 Partner	-0.01	0.72	-0.01	0.72	-0.02	0.72
5	Understand	GPT-3 Partner	-0.02	0.23	-0.03	0.23	-0.04	0.23
6	Respect	GPT-3 Partner	-0.02	0.23	-0.03	0.23	-0.04	0.23
7	Divisiveness Index	GPT-3 Self	-0.01	0.70	-0.01	0.70	-0.01	0.70
8	Affective Polarization	GPT-3 Self	-0.01	0.72	-0.01	0.72	-0.01	0.72
9	Good Reasons	GPT-3 Self	-0.01	0.51	-0.02	0.51	-0.02	0.51
10	Point of View	GPT-3 Self	-0.01	0.79	-0.01	0.79	-0.01	0.79
11	Understand	GPT-3 Self	0.01	0.77	0.01	0.77	0.01	0.77
12	Respect	GPT-3 Self	-0.01	0.73	-0.01	0.73	-0.01	0.73

Table 22: Divisiveness Index TSLS CACE Analysis: Cell values are average marginal effects (coefficients) and p-values calculated from a two-stage least squares CACE model, with standard errors clustered by conversation. Instruments are treatment assignment and a binary indicator for whether the partner sent a single message. Controls are pre-chat position on gun control, party ID, employment, education, and race.

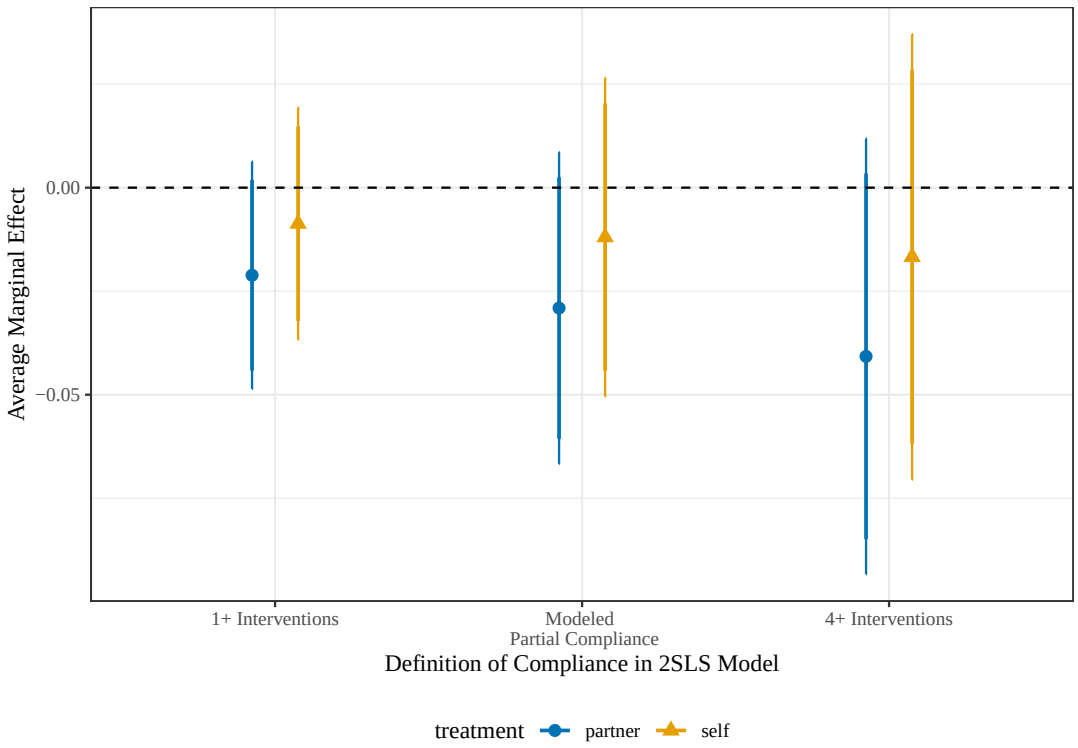

Figure 17: Alternative Divisiveness TSLS CACE results, with no controls. This basic model includes only the treatment assignment as an instrument for treatment dosage.

H.3 Change in Gun Policy Attitudes

Figure 18 presents estimated treatment effects on the absolute value of the change in an index of gun policy positions from before the conversation to after the conversation. Because we do not expect the conversations, which are balanced between matched pairs of gun control supporters and opponents, to result in general movement to the right or the left, we examine the absolute value of change in responses. Lower values indicate less movement in attitudes. As Figure 18 illustrates, we find no evidence of treatment effects on individuals' post-chat gun control policy positions. GPT-3 rephrasings do not seem to increase persuadability on the issues, even while they make the people and arguments on the other side appear more reasonable. This provides evidence that these kinds of AI tools can be used to improve divisive political conversations without manipulating respondents to adopt a particular political viewpoint.

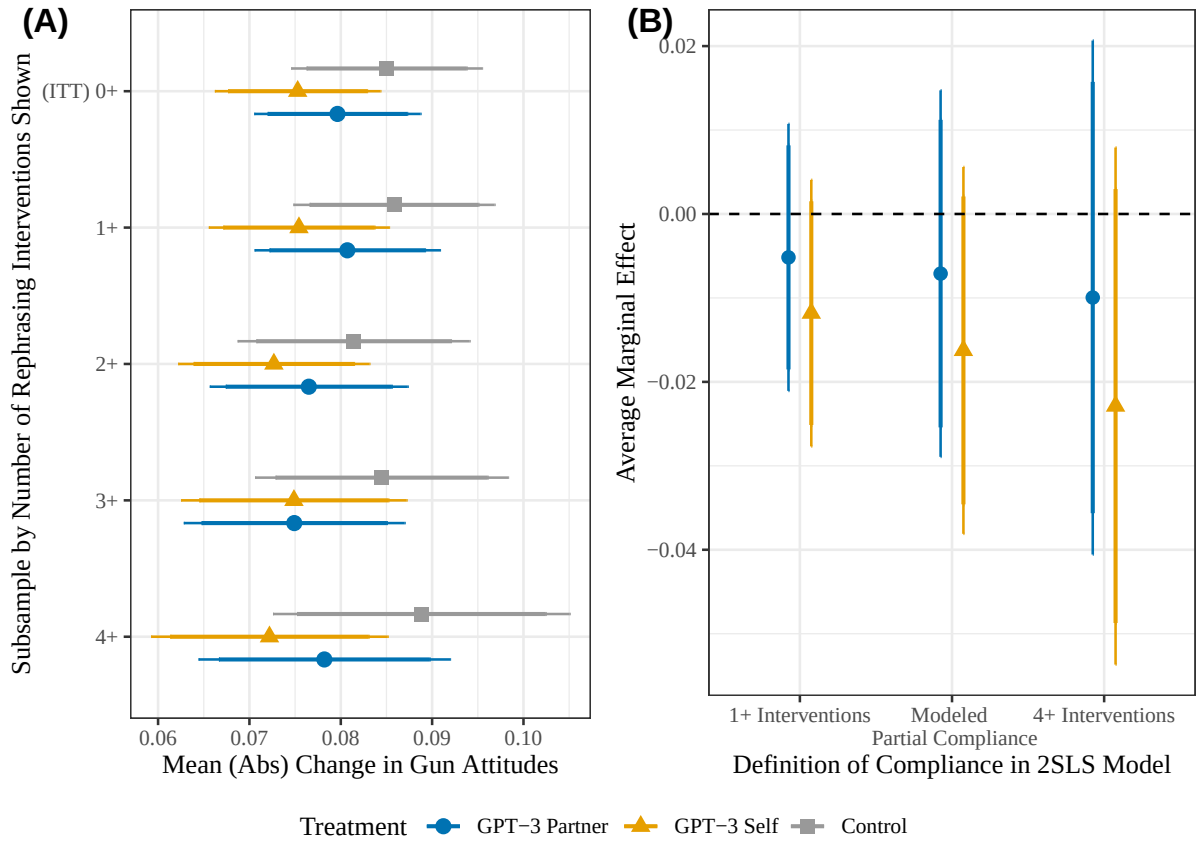

Figure 18: Analysis of Change in Gun Policy Attitudes. The index is scaled from 0 (no change in attitude) to 1 (complete reversal of opinions), and is the absolute value of the post-survey value minus the pre-survey value. The number of rephrasings are overlapping sets, such that 0+ includes all observations. The left panel presents the means and 95% confidence intervals based on unadjusted standard errors. The right panel shows average marginal effects from the CACE analysis, with standard errors clustered at the conversation level. There is no evidence of the treatment increasing attitude change.

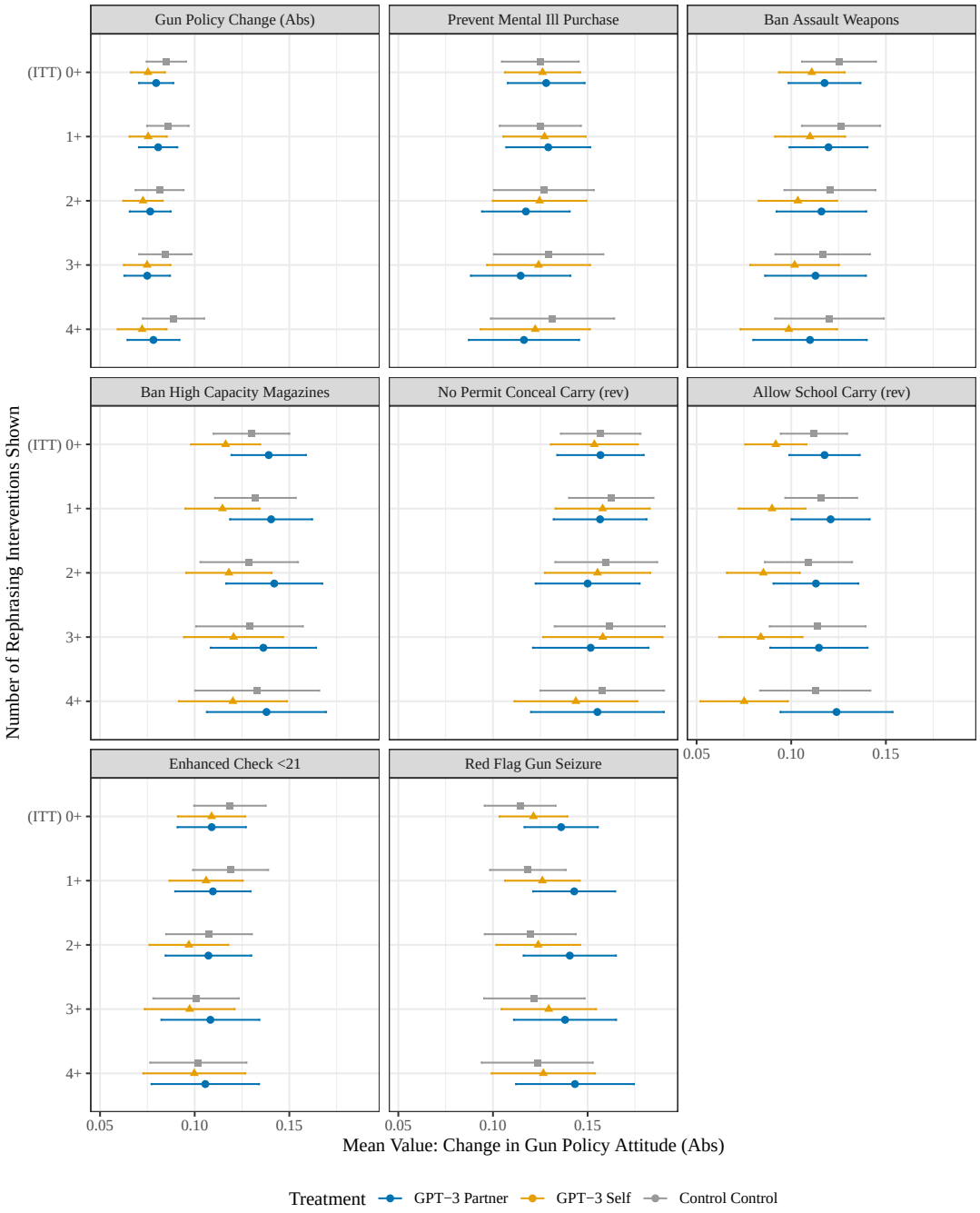

Figure 19: Mean and 95% confidence interval for all metrics that make up the gun policy index, by treatment condition and conversation length. Gun policy attitudes are the absolute value of the change in opinion between the pre-survey and the post-survey.

	DV	Treatment	0+ mean	0+ sd	1+ mean	1+ sd	2+ mean	2+ sd	3+ mean	3+ sd	4+ mean	4+ sd
1	Gun Policy Change (Abs)	Control	0.09	0.12	0.09	0.12	0.08	0.12	0.08	0.12	0.09	0.13
2	Gun Policy Change (Abs)	Partner	0.08	0.11	0.08	0.11	0.08	0.10	0.07	0.10	0.08	0.11
3	Gun Policy Change (Abs)	Self	0.08	0.11	0.08	0.11	0.07	0.10	0.07	0.10	0.07	0.10
4	Prevent Mental Ill Purchase	Control	0.12	0.24	0.12	0.24	0.13	0.24	0.13	0.25	0.13	0.25
5	Prevent Mental Ill Purchase	Partner	0.13	0.24	0.13	0.24	0.12	0.22	0.11	0.22	0.12	0.23
6	Prevent Mental Ill Purchase	Self	0.13	0.23	0.13	0.23	0.12	0.23	0.12	0.23	0.12	0.22
7	Ban Assault Weapons	Control	0.13	0.23	0.13	0.23	0.12	0.22	0.12	0.21	0.12	0.22
8	Ban Assault Weapons	Partner	0.12	0.22	0.12	0.22	0.12	0.22	0.11	0.23	0.11	0.23
9	Ban Assault Weapons	Self	0.11	0.20	0.11	0.20	0.10	0.20	0.10	0.20	0.10	0.20
10	Ban High Capacity Magazines	Control	0.13	0.23	0.13	0.24	0.13	0.24	0.13	0.24	0.13	0.25
11	Ban High Capacity Magazines	Partner	0.14	0.23	0.14	0.23	0.14	0.24	0.14	0.24	0.14	0.24
12	Ban High Capacity Magazines	Self	0.12	0.21	0.11	0.21	0.12	0.21	0.12	0.22	0.12	0.22
13	No Permit Conceal Carry (rev)	Control	0.16	0.25	0.16	0.25	0.16	0.25	0.16	0.25	0.16	0.25
14	No Permit Conceal Carry (rev)	Partner	0.16	0.27	0.16	0.26	0.15	0.26	0.15	0.26	0.16	0.27
15	No Permit Conceal Carry (rev)	Self	0.15	0.27	0.16	0.27	0.16	0.26	0.16	0.27	0.14	0.25
16	Allow School Carry (rev)	Control	0.11	0.21	0.12	0.21	0.11	0.21	0.11	0.22	0.11	0.23
17	Allow School Carry (rev)	Partner	0.12	0.22	0.12	0.22	0.11	0.21	0.11	0.22	0.12	0.23
18	Allow School Carry (rev)	Self	0.09	0.19	0.09	0.19	0.09	0.18	0.08	0.19	0.08	0.18
19	Enhanced Check <21	Control	0.12	0.22	0.12	0.22	0.11	0.21	0.10	0.19	0.10	0.20
20	Enhanced Check <21	Partner	0.11	0.21	0.11	0.21	0.11	0.21	0.11	0.22	0.11	0.22
21	Enhanced Check <21	Self	0.11	0.21	0.11	0.21	0.10	0.20	0.10	0.20	0.10	0.21
22	Red Flag Gun Seizure	Control	0.11	0.22	0.12	0.22	0.12	0.22	0.12	0.23	0.12	0.23
23	Red Flag Gun Seizure	Partner	0.14	0.23	0.14	0.23	0.14	0.23	0.14	0.23	0.14	0.24
24	Red Flag Gun Seizure	Self	0.12	0.21	0.13	0.21	0.12	0.21	0.13	0.21	0.13	0.21

Table 23: Gun Policy Opinion Index Means: Columns indicate the mean and standard deviation for the gun policy index, followed by each constituent item. The values are the mean of the absolute value of the post-treatment position minus the pre-treatment position. Results shown by subsets of the minimum number of rephrasings that was shown to the treated respondent (or that would have been shown in a control condition if it had been treated.)

H.4 Over Time Results

As noted in the main text, we sent a follow-up survey to the original study participants approximately 3 months after the chat experiment. We asked respondents to answer the same questions to measure divisiveness and gun policies as in the study presented here. To evaluate the persistence of these effects, we re-ran the main treatment analyses with these follow-up measures. We estimated these only for the divisiveness index as we did not observe effects on the gun policy items during the original experiment and we did not ask the conversation quality questions on the follow-up (as it made little sense with the passage of time to do so). Figure 20 displays these results, showing no evidence of persistence of the key effects noted in our study. Given the fleeting, one-time nature of our treatment, this is not surprising.

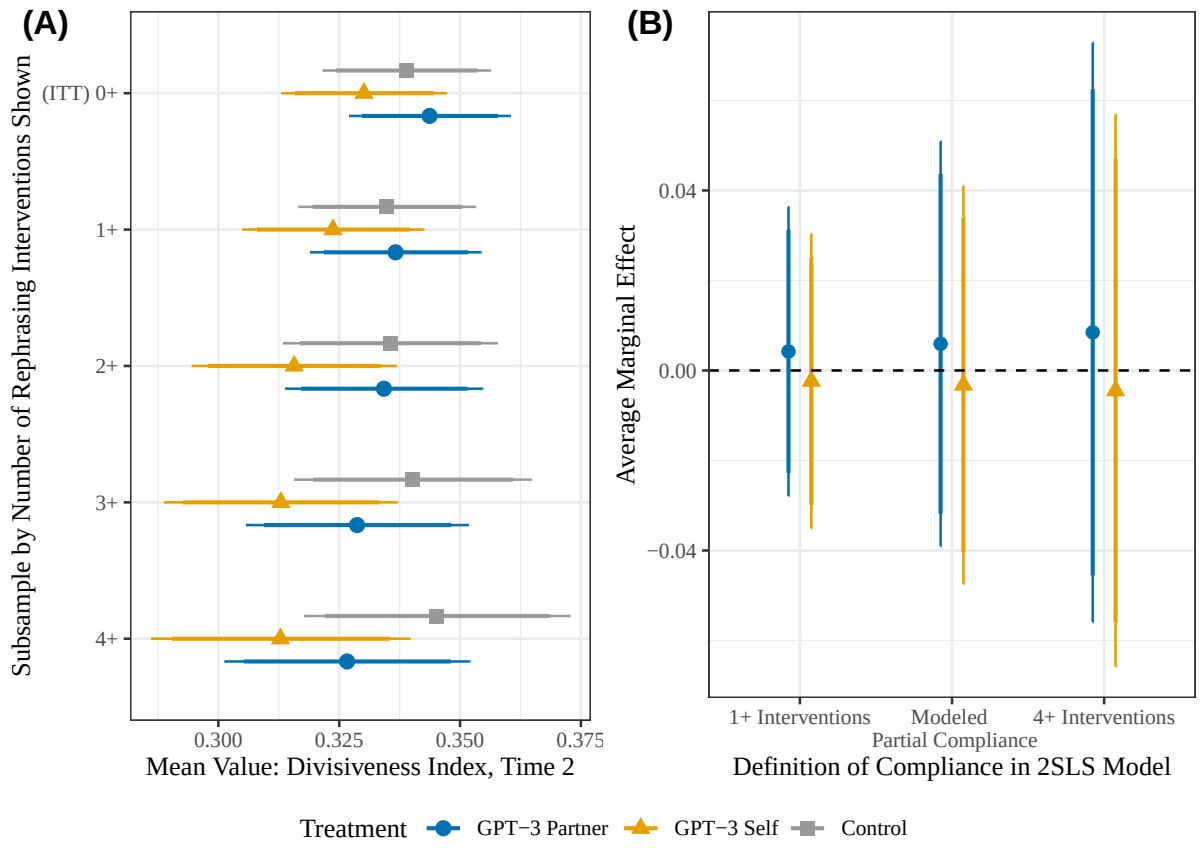

Figure 20: Analysis of divisiveness index on the follow-up survey. The index is scaled from 0 (lowest level of divisiveness) to 1 (highest divisiveness on all conversation measures). The number of rephrasings are overlapping sets, such that 0+ includes all observations. The left panel presents the means, 90% and 95% confidence intervals based on unadjusted standard errors. The right panel shows average marginal CACE effects with standard errors clustered at the conversation level. We observe no treatment effects approximately 3 months after the original experiment.