Appendices for "Introducing ReChat: A Lab-in-the-Cloud for Text Discussions"

Xiaoxiao Shen and William Small Schulz

Appendix 1. ReChat Template Variables

Table 3. ReChat Template Variables

Variable	Values	Uses
Room Template Name	Text	Name template for internal (researcher) use
Description	Text	Describe template for internal (researcher) use
Public Name	Text	Specify heading in chat window seen by participants
Chat End Mode	By Time By Questions	Determines whether the chat ends after the Expiration Time, or after all Questions have been posed
Expiration Time	Minutes	Set maximum duration of chat, when Chat End Mode is set to By Time
Chat Questions	Type, Timing, Content, and Response Options	Pose multiple-choice questions or give simple instructions at set time intervals during the chat (when Chat End Mode is set to By Questions, the chat ends after the last question/prompt is posed/given)
Waiting Time	Minutes	Determine how long participants can be kept waiting in the waiting room
Waiting Room Timeout Message	Text	Deliver an apologetic message to participants who timed out in the waiting room
Number of Participants	Number	The number of participants per room
End Instruction	Text	Deliver a thankful/instructional message to participants at the conclusion of the chat
Prefix for completion code	Text	A prefix applied to the random string generated as the participant's completion code, that indicates they successfully completed a chat
Prefix for timeout completion code	Text	A prefix applied to the random string generated as the participant's completion code, that indicates they timed out in the waiting room

Appendix 2. Selection of Hypocrisy as Trait for Discussion

In designing our discussion stimuli, we hoped to give respondents an opportunity to express as much partisan affect they like, without constraining the topic of discussion unnecessarily. An early pilot asked participants to compare the intelligence of democrats and republicans, and many chat participants asserted that they did not think there was a difference, and some stated that they considered this to be a bad question in the first place. So, we sought to identify a trait on which partisans would be willing to express more meaningful in-group favoritism.

To do this, we re-analyzed data from Druckman et al. 2022: Figure 5 displays rates of in-group favoritism in trait ratings originally collected by Druckman et al. 2022. Traits are adjectives, listed along the x axis, and respondents were asked, "How well does '[trait]' describe [Democrats/Republicans]" on a 5-point likert scale. To produce this plot, the difference was taken between partisans' ratings of out-partisans and their ratings of in-partisans, recoded so that all differences reflected in-group favoritism (in effect, responses for negative traits were flipped), and binarized such that cases of in-group favoritism (in-group rated 'better' than out-group) were coded as 1, and 0 else (n.b. for all traits, the modal response was to rate the in-group and out-group equally, and the median response was to rate the in-group one point 'better'). The mean of this binarized variable was calculated for all partisans (black), Democrats (blue), and Republicans (red). This summarizes aggregate propensity to express in-group favoritism on each trait, in each of these subsets of the sample.

Inspecting the black points, we can see that partisans overall are most likely to express in-group favoritism with respect to open-mindedness, generosity, and hypocrisy. However, open-mindedness and generosity also display the largest inter-party differences, which is undesirable if we are seeking to develop a prompt that can be used in studies of both Democrats and Republicans (Republicans were not the focus of the present studies, but we plan to include them in future studies). Furthermore, it is reasonable to suspect that there actually are substantive partisan differences in open-mindedness and generosity, with respect to specific social and economic issue preferences, and this is incompatible with our goal of giving participants a generic outlet for their partisan affect. We therefore selected hypocrisy as the focus of our prompt in these studies, since it has the smallest inter-party difference in in-group favoritism, while still having the third-highest rate of in-group favoritism overall. Hypocrisy is also the only trait for which both parties tend to favor the in-group overall. Finally, it is theoretically possible to make a reasonable argument that either party is hypocritical, with respect

to any given issue, and thus satisfies our goal of giving respondents an open-ended opportunity to express their partisanship.

Several ethical considerations attend our choice of a prompt that encourages discussion of partisanship. In particular, we anticipated that this discussion might increase participants' affective polarization, which is generally considered detrimental to democratic norms. However, for this very reason, we believe it is important to study conversational dynamics that may contribute to this phenomenon. Moreover, we consider that the dosage of our treatment was well within the normal range of what participants might encounter in their daily life, and we expected any effects of this treatment on partisan affect would fade over time, consistent with evidence from similar designs (e.g. Santoro and Broockman 2022). It should also be noted (as discussed further in Appendix 11) that the magnitude of the increase in affective polarization we observed was considerably smaller than that observed by Broockman, Kalla, and Westwood (2023), who implemented treatments designed to induce the largest possible differences in partisan affect.

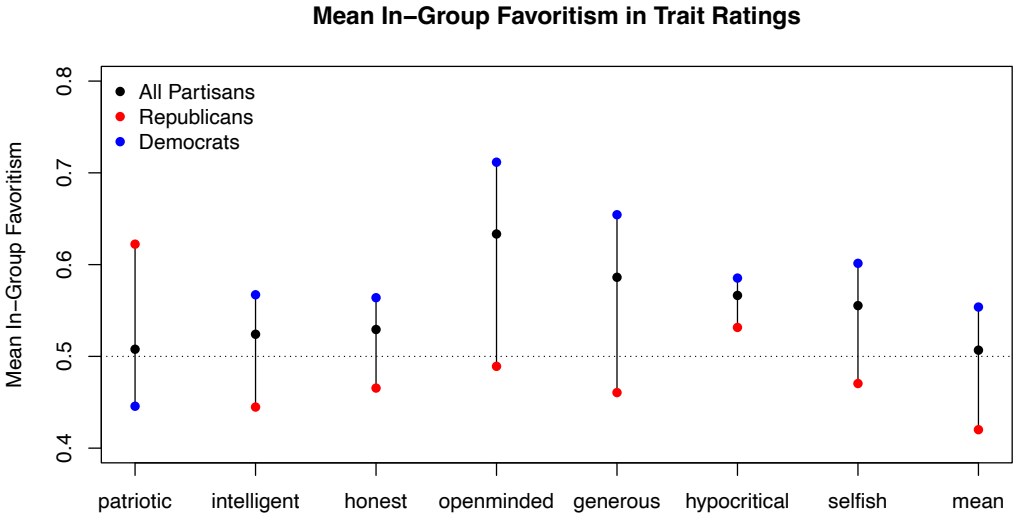


Figure 5. Mean favoritism for in-group over out-group in ratings of 8 traits. Red points represent Republicans' probability of rating Republicans more favorably than Democrats. Blue points represent Democrats' probability of rating Democrats more favorably than Republicans. Black points represent probability of rating co-partisans more favorably than out-partisans in the pooled sample (that is, amongst all Democrats and Republicans in the original survey sample, excluding only pure independents.

Appendix 3. Distribution of Partisanship Strength

Because all participants in our analyses were Democrats, our 6-point measure of partisanship is a *de facto* inverted 3-point measure of partisanship strength. So, we include such a measure in our models as "partisanship strength." Figure 6 displays the distribution of this variable in the pooled Study 1 and Study 2 data.

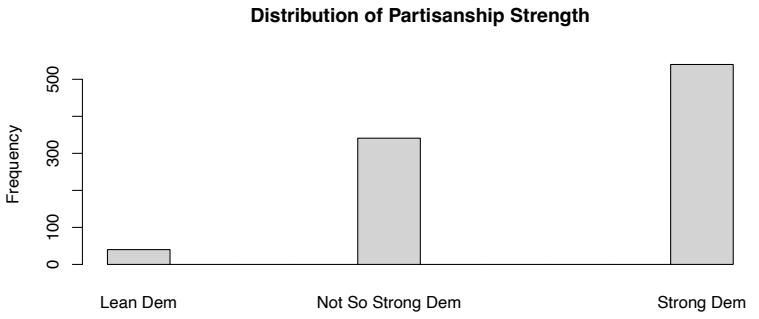


Figure 6. Distribution of 3-point partisanship strength measure.

Appendix 4. Recruitment Language

This appendix provides language used in posting our Human Intelligence Tasks (HITs) on Amazon Mechanical Turk. A HIT requires a **title**, and may optionally include a **description**.

Appendix 4.1 Study 1

HIT Title: "Survey (up to $2.75 including potential bonuses!)"

HIT Description: "Take a short survey for $0.75, and optionally participate in a 5-minute activity for bonus of up to $2 (up to $2.75 total)"

Appendix 4.2 Study 2

Note: Recruitment for Study 2 begin similarly to that for Study 1, in that the HIT Title and Description withheld information about the chat activity. However, this approach was costly from the perspective of analyzing chat outcomes, since it required recruiting participants to take the survey who would not be interested in participating in a chat. We therefore revised the HIT Title and Description to provide details about the chat, since we determined that Study 1 was sufficient for analyzing chat self-selection (which was the main purpose for which we had withheld information about the chat from initial recruitment). Overall, 303 individuals took the Study 2 HIT under the original (chat-withheld) advertising, and 336 took the Study 2 HIT under the revised (chat-declared) advertising.

HIT Title (1): "Survey (up to $3.75 including potential bonuses!)"

HIT Description (1): "Take a short survey (approximately 5 minutes) for $0.75, and optionally participate in a 10-minute activity for bonus of up to $3 (up to $3.75 total)"

HIT Title (2): "Political chat with a fellow Democrat (up to $3.75 including potential bonuses!)"

HIT Description (2): "Take a short survey (under 5 minutes) for $0.75, and for a $2 bonus participate in a 10-minute political chat with another Democrat. Additional $1 bonuses granted if your thoroughness in the chat is rated above the median average (so half of all participants will receive this additional thoroughness bonus, for total compensation of $3.75)."

We note that in Study 2, advertising the activity as a "political chat" might have affected how participants engaged with the study, along the lines suggested by Groenendyk and Krupnikov (2021).

Appendix 5. Predictors of Extroversion

This appendix provides a supplementary analysis of several possible predictors of extroversion, in a linear regression framework.

Table 4. Predictors of Extroversion

	Dependent variable:
	Extroversion
Political Interest	0.202*** (0.043)
College	0.192 (0.170)
Social Media Expressor	0.220** (0.095)
Constant	–1.098*** (0.224)
Observations	483
R^2	0.069
Adjusted R^2	0.063
Residual Std. Error	0.968 (df = 479)
Note:	*p<0.1; **p<0.05; ***p<0.01

Appendix 6. Study 1 Free-Text Explanations

These are some representative examples of free-text explanations offered by participants in Study 1, regarding why they did not wish to participate in a chat. We offer this as qualitative evidence that abstention from chat participation is associated with introversion or identifying as "shy."'

> "First of all I am shy and also discussions about politics never yield any results."
>
> "I worry about being able to carry the conversation well, and the anxiety was not worth the potential bonus to me."
>
> "I'm a shy person ... I don't feel like I have anything insightful to add to a conversation."
>
> "I just really don't want to interact with others. I don't want to talk about politics either, its such a drag now."

The online supplementary materials include replication code for a Structural Topic Model (Roberts et al. 2014) that analyzes these free-responses quantitatively.

Appendix 7. Study 2 Recruitment and Attrition

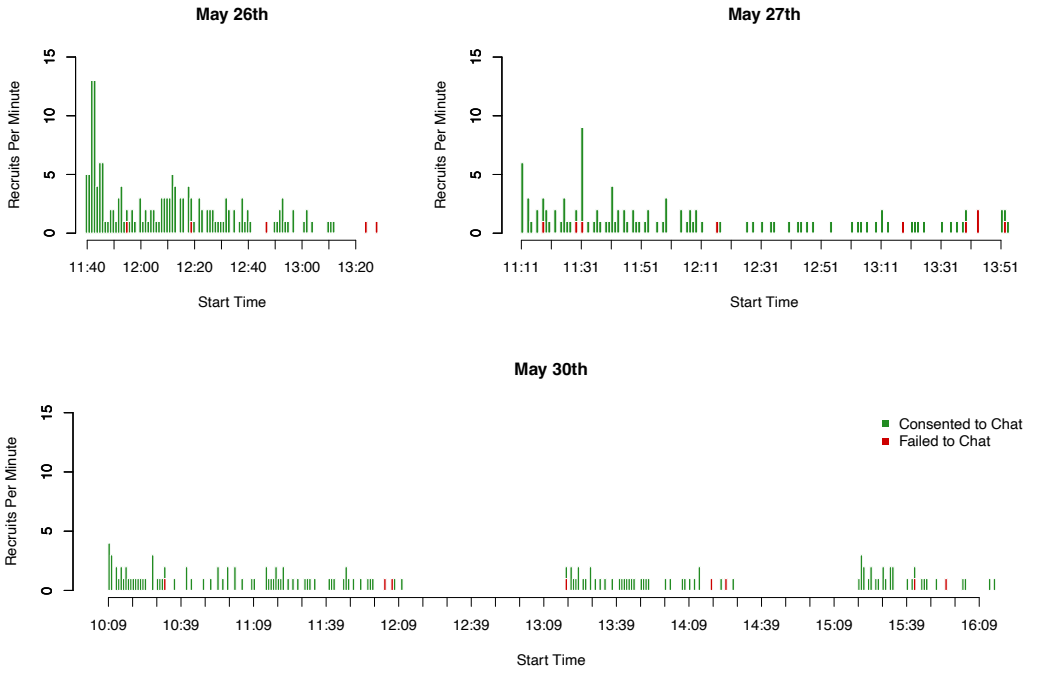


Figure 7. Study 2 recruitment, conducted over three days.

Figure 7 displays recruitment over time for Study 2, in the same way Figure 3 displays this information for Study 1, with one important distinction: when Study 2 began advertising the chat activity in its MTurk description, we could no longer analyze agreement to take the survey as distinct from agreement to participate in a chat. So, we drop the "Took Survey" category that was included in visualizing Study 1 recruitment.

Recruitment for Study 2 was spread out over three days: May 26th, 27th, and 30th, 2022. We stopped recruitment on the first day (the 26th) after approximately 2pm EDT, with the goal of focusing recruitment during the hours we expected MTurkers to be the most active. We stopped recruitment on the second day (the 27th) slightly later. We did not conduct recruitment on the 28th or 29th because this was a weekend. We therefore resumed recruitment on May 30th. Recruitment was slow on May 30th, and we halted recruitment before reaching our target sample because we had used more funds on Day 1 than intended (we later noted that May 30th was Memorial Day, which might also explain the slowness). We acknowledge that this early stopping represents an issue in terms

of researcher-degrees-of-freedom, however we did anticipate this possibility in our pre-registration plan, and took the decision without knowing the results of our hypothesis tests (although we did monitor some aspects of data collection, such as completion rate and refusal rate, in real time). As such, we believe the data we collected is more than sufficient for demonstration purposes.

We did find some evidence that failure-to-chat was associated with the accountability treatment in Study 2 (p = .024), which may help to explain the imbalance in treatment assignment in the data we ultimately analyzed (44% of participants who completed the chat were assigned to the neutral accountability condition). Although we do not wish to over-interpret this finding, it reinforces the need for analysts of chat experiments to carefully consider how self-selection factors shape the sample of participants who generate their data.

Additionally, in light of the above-noted association between the accountability treatment and attrition, we estimated Lee (2009) bounds on the accountability treatment effect, using an R procedure adapted from the STATA procedure recommended by Tauchmann (2014). Figure 8 displays the Lee bounds (×'s) for the accountability treatment coefficient estimated for each outcome in Table 8 (dots). In all four cases, the bounds contain zero, which is to be expected, considering that the main versions of these analyses found relatively weak evidence for a treatment effect on these outcomes.

To avoid problems like these, future researchers should remember that it is generally preferable to apply the treatment as late in the study process as is feasible. For example, a revised version of the present design could apply the treatment through an informational message supplied by the moderator-bot, so that participants are already present in the chat (which may reduce their attrition). In general, it is also expected that increasing monetary compensation will reduce attrition (and thus, imbalance) as well.

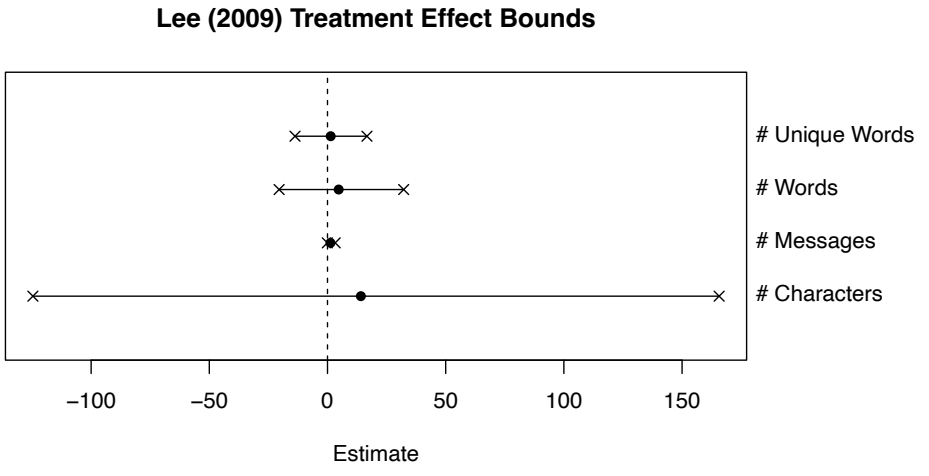


Figure 8. Lee (2009) bounds on accountability treatment effect for all four outcome variables.

Appendix 8. Mobile versus Desktop

We conducted supplementary analyses to assess whether there were important differences between participants who took our studies on mobile versus desktop devices. To do this, we used information about the browser the participant used when taking the Qualtrics survey, and categorized the browsers according to whether they were typically used on mobile or desktop devices. Approximately 6% of participants appeared to have used a mobile device for the study.

We analyze the relationship between mobile usage and several variables we considered potentially relevant. First, we analyzed whether participating via mobile was associated with reduced loquaciousness, as might be expected given the greater difficulty of typing extensively on a mobile device, compared to a computer keyboard. Indeed, Table 5 shows that mobile usage was associated with significantly lower loquaciousness by all four metrics.

Second, we analyzed whether taking the study on mobile was associated with participants experiencing technical issues with the chat. Table 6 offers no evidence that user-reported technical issues were more prevalent amongst mobile users; however, we found that participants' age was a strongly significant predictor of technical issues. This suggests that future implementers of chat studies should bear in mind that their sample may be filtered to some extent on subjects' digital literacy (although this filtering is plausibly externally-valid to real-world text-based conversations).

Table 5. Mobile Devices and Loquaciousness

	Dependent variable:			
	char_count	message_count	word_count	unique_word_count
	(1)	(2)	(3)	(4)
took_on_mobile	–161.861*** (51.443)	–1.275* (0.663)	–28.744*** (9.382)	–19.156*** (5.473)
as.factor(study)2	512.137*** (24.445)	4.906*** (0.315)	94.939*** (4.458)	57.778*** (2.601)
Constant	394.692*** (18.801)	4.235*** (0.242)	70.709*** (3.429)	54.817*** (2.000)
Observations	669	669	669	669
R^2	0.400	0.268	0.408	0.429
Adjusted R^2	0.398	0.266	0.406	0.427
Residual Std. Error (df = 666)	311.481	4.014	56.808	33.136

Note: *p<0.1; **p<0.05; ***p<0.01

Table 6. Mobile Devices and Technical Problems Reported

	Dependent variable:
	!is.na(had_problem_1)
took_on_mobile	0.022 (0.046)
as.factor(study)2	0.004 (0.023)
age	0.005*** (0.001)
Constant	–0.075* (0.038)
Observations	747
R^2	0.035
Adjusted R^2	0.031
Residual Std. Error	0.306 (df = 743)
F Statistic	8.926*** (df = 3; 743)
Note:	*p<0.1; **p<0.05; ***p<0.01

Appendix 9. Theorizing and Operationalizing Loquaciousness

Loquaciousness, in general, refers to the quality of being very talkative or chatty. In the context of this study, we were particularly interested in loquaciousness in the sense of "saying a lot," such as would be expected to be associated with conveying a large number of ideas, providing extensive evidence or reasoning to back up one's arguments, having a large persuasive influence over their audience's subsequent views (or possibly a normative influence over the views their audience perceives as socially-valid), and generally exercising social dominance in a conversation.

These qualities have previously been studied in face-to-face conversations. For example, Karpowitz and Mendelberg (2014) studied various aspects of authority in deliberative settings, using measures such as floor time (the amount of time participants spend speaking during the interaction) and the types and frequency of interruptions. However, neither of these measures translates easily to conversations that take place via text-based chats, because there is not a clear (to the participants or to the researcher) amount of time that a chat participant holds the "floor," and (as a consequence of this) it is difficult to say whether any given message that interposes between two other messages from another participant constitutes an "interruption."

To transition from traditional to digital contexts, this paper operationalizes loquaciousness as the total textual output of each study participant in an online chat. As discussed in the following sections, we consider character count, message count, word count, and unique word count, but we pre-register the Study 2 analyses with respect to character count. This straightforward approach acknowledges the textual nature of online communication as the primary medium for interaction, captures the extent of a participant's engagement and communicative output in a quantifiable manner, and provides a transparent, objective measure that can be easily replicated across studies, enhancing the reproducibility of research findings.

Appendix 10. Study 1 Loquaciousness Analyses

We conducted an exploratory analysis of loquaciousness in Study 1, which we used to guide Study 2. We constructed several measures of loquaciousness for each participant: number of messages sent in the chat, total character count, total word count, and total unique word count. Table 7 presents results with each of these metrics as the dependent variable. These analyses are based on linear regression with robust standard errors clustered at the level of the chatroom.

First, for all metrics apart from message count, we observed a significant correlation between loquaciousness and ideology: liberals used significantly more characters, words, and unique words than moderates.

Furthermore, we observe a significant gender effect: compared to participants who identified their gender as female or non-binary, those who identified as male sent .6 more messages, used 69 more characters, and said 13 more words and 9 more unique words.

Finally, we observe a treatment effect: compared to co-partisan accountability-treated participants, on average neutral-treated participants sent .5 more messages, using 56 more characters, 10 more words, and 7 more unique words. This suggests that having one's thoroughness evaluated by a neutral party induced significantly greater thoroughness than being evaluated by co-partisans.

Table 7. Loquaciousness (Study 1)

	Dependent variable:			
	Character Count	Message Count	Word Count	Unique Word Count
	(1)	(2)	(3)	(4)
Treatment				
(Neutral Accountability)	56.413** (22.523)	0.501* (0.259)	10.278** (4.118)	6.854** (2.709)
Extroversion	–9.678 (12.437)	–0.154 (0.153)	–1.745 (2.275)	–1.081 (1.551)
Self-Monitoring	–6.319 (15.074)	0.104 (0.145)	–1.327 (2.671)	–1.349 (1.682)
Political Interest	17.573 (12.196)	0.136 (0.135)	2.826 (2.205)	1.870 (1.437)
Partisanship Strength	–9.183 (23.768)	0.044 (0.351)	–0.607 (4.229)	–1.249 (2.847)
Affective Polarization	0.410 (0.625)	0.003 (0.006)	0.058 (0.111)	0.056 (0.074)
Ideological Extremity	36.944** (14.538)	–0.096 (0.203)	5.613** (2.677)	3.805** (1.768)
Ideological Identity Strength	–19.982 (14.327)	0.032 (0.210)	–3.031 (2.647)	–2.188 (1.830)
Media Consumption Scale	–0.267 (10.210)	–0.101 (0.137)	–0.415 (1.855)	–0.310 (1.220)
Male	68.950*** (23.746)	0.618** (0.277)	13.166*** (4.378)	9.470*** (2.897)
Age	–0.976 (0.970)	–0.016 (0.012)	–0.156 (0.180)	–0.074 (0.120)
College	–5.752 (32.555)	0.019 (0.513)	–2.231 (6.121)	–1.903 (3.964)
Social Media Expressor	19.153 (24.141)	0.046 (0.279)	3.996 (4.353)	2.252 (2.929)
Constant	210.848*** (73.405)	3.609*** (0.905)	39.817*** (13.790)	33.813*** (9.113)
Observations	278	278	278	278
R^2	0.113	0.062	0.103	0.115
Adjusted R^2	0.069	0.015	0.059	0.072
Residual Std. Error (df = 264)	188.152	2.111	34.377	22.315

Note: *p<0.1; **p<0.05; ***p<0.01

Appendix 11. Study 2 Detailed Analyses

Study 2 sought to conduct a pre-registered replication of the findings of the exploratory analyses of loquaciousness in Study 1 (presented in Appendix 10). We doubled the chat duration to 10 minutes, to allow for more variance in loquaciousness, and we pre-registered[9] three hypotheses:

H1 **Males are more loquacious than those who identify as female or other gender.** Prior studies (e.g. Mendelberg, Karpowitz, and Oliphant 2014) found women often have less influence than men, in certain face-to-face deliberative settings. So, we were interested in whether gender differences in loquaciousness would arise in the ReChat environment.

H2 **Stronger ideologues[10] are more loquacious than those whose ideology is more moderate.** Differential loquaciousness has been proposed as an explanation for the apparent polarization of online political discourse: for example, Hughes (2019) found that 73% of tweets about US national politics come from a small group of ideologically-extreme users, and Bail (2021) described the absence of moderate voices as "the most profound form of distortion" (p. 82) in online discourse. So, we sought to test whether this distortion could be reproduced in the ReChat environment.

H3 **People in the neutral "thoroughness" accountability treatment condition are more loquacious than those in the Democratic (co-partisan) accountability treatment condition.** As noted above, this treatment was intended to simulate different "imagined audiences," (Marwick and boyd 2011) and we expected that participants would anticipate a co-partisan audience would hold them to a lower standard of "thoroughness" than a neutral audience would.

These hypotheses were consistent with exploratory analyses of the Study 1 data, in that male gender was significantly associated with sending more *messages*, containing more *characters*, more *words*, and more *unique words*; more extreme (liberal) ideology was associated with writing more characters, words, and unique words; and the neutral treatment was associated with greater loquaciousness by all four metrics. In Study 2, we conducted a pre-registered replication of these findings.

To limit researcher-degrees-of-freedom, we pre-registered our planned analyses with respect to *character count*. Character count was chosen because it is a simple, objective, and quantifiable summary of participants' communicative output. Although it does not necessarily differentiate quality from mere verbosity, it does provide an indicator of participants' effort, and it is easy to interpret, which

9. https://osf.io/atfyq
10. Note that in this all-Democrat sample, this is equivalent to testing whether liberals are more loquacious than moderates.

makes subsequent analyses more transparent.

Additionally, we conducted exploratory analyses using the pre-post feeling thermometers, to address the following research questions:

RQ1 **Group Polarization:** Is post-chat partisan affect more polarized than pre-chat partisan affect?

RQ2 **Causes of Polarization:** What factors explain variation in post-chat partisan affect?

Appendix 11.1 Results

Appendix 11.1.1 Planned Analyses

Our planned analyses regressed total character count[11] (Mean = 896, SD = 376) on the three explanatory variables – an indicator for male gender, an ordinal self-reported ideology measure, and an indicator for receiving the "neutral thoroughness judge". We set a threshold of $\alpha = .05$ in a two-sided test, with standard errors clustered at the chatroom level. Analyses of Study 1 data indicated that a sample of N=500 would achieve 80% power in our planned analyses, although for budgetary reasons we halted recruitment earlier, such that N=390 cases were ultimately included (see Appendix 7 for a detailed discussion of attrition) from 207[12] chatrooms. Simple random assignment placed 171 subjects in the "Neutral" judge treatment, and 219 in the "Democratic" judge treatment.

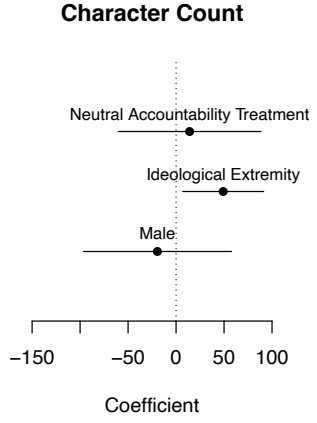
The planned analyses (see Table 8 Column 1, and Figure 9a) rejected the null for Hypothesis 2: within this all-Democrat sample, ideological extremity was significantly associated with greater loquaciousness ($p = 0.023$). We did not reject the null for Hypotheses 1 and 3: neither the coefficient on male gender ($p = 0.622$) nor the treatment coefficient ($p = 0.710$) were significantly different from zero at the planned thresholds, though the treatment coefficient did have the expected sign. Exploratory analyses lent some nuance to these findings: when considering *message* count as the dependent variable, we rejected the null for Hypotheses 1 ($p = 0.028$) and 3 ($p = 0.018$), as shown in Table 8, column 2, and Figure 9b. We consider this suggestive evidence that gender and the accountability treatment are related to loquaciousness, albeit less robustly than is ideological extremity.

11. We chose to operationalize loquaciousness as character count because it had a roughly linear relationship with word count and unique word count, and was simpler to interpret.

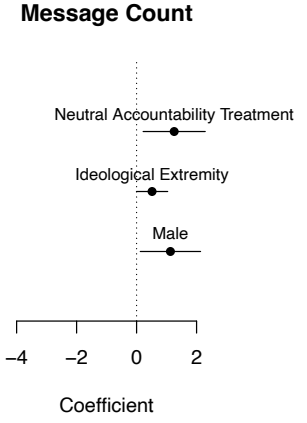
12. This number is greater than half the number of cases, because some cases' partners either failed to enter their chat completion code or failed to answer one of the survey questions used in this analysis, and so had to be dropped.

Table 8. Loquaciousness (Study 2)

	Dependent variable:			
	# Characters (Pre-Reg)	# Messages	# Words	# Unique Words
	(1)	(2)	(3)	(4)
Treatment				
(Neutral Accountability)	14.124 (37.953)	1.253** (0.528)	4.730 (7.018)	1.350 (3.985)
Ideological Extremity	49.128** (21.520)	0.516* (0.266)	8.588** (3.922)	5.887*** (2.249)
Male	–19.456 (39.386)	1.128** (0.513)	–5.114 (7.155)	–1.284 (4.128)
Constant	803.724*** (51.802)	7.022*** (0.610)	147.349*** (9.485)	99.950*** (5.484)
Observations	390	390	390	390
R^2	0.012	0.033	0.013	0.015
Adjusted R^2	0.004	0.025	0.005	0.007
Residual Std. Error (df = 386)	375.307	4.888	68.442	39.000
F Statistic (df = 3; 386)	1.574	4.364***	1.667	1.965
Note:				*p<0.1; **p<0.05; ***p<0.01



(a) Character count as DV (pre-registered).



(b) Message count as DV (exploratory).

Figure 9. Predictors of loquaciousness during chats in Study 2, plotted as linear regression coefficients from the pre-registered model with raw character count as the dependent variable (Figure 9a), and a planned exploratory analysis with message count as the dependent variable (Figure 9b). Note that in this sample, ideological extremity is equivalent to greater liberalism. Whiskers show 95% confidence intervals.

Appendix 11.1.2 Exploratory Analyses

Exploring the possibility that extremists' loquaciousness might make these conversations particularly polarizing, we found that the difference in feeling thermometer ratings of Democrats and Republicans was on average 2.9 points higher after the chat than before the chat ($p < 0.001$, 2-tailed paired t-test), which suggests[13] that chats polarized participants' partisan affect. To illustrate the magnitude of this difference by comparison to another study that sought to manipulate affective polarization, Broockman, Kalla, and Westwood (2023) report a pair of treatments that (together) produced a 14 point difference in respondents' affective polarization (measured in the same way), similar to the magnitude by which affective polarization increased in the United States between 1978 and 2008. However, it should be noted that the change observed in the present study occurred following a relatively brief interaction with a stranger, so if individuals are exposed to political conversations like these on a regular basis, there may be large cumulative consequences.

Table 10 (column 2) shows evidence consistent with persuasion: baseline affect had a significant effect[14] on partner endline affect. However, simple persuasion cannot explain aggregate polarization, since if partners converged to each others' priors, aggregate change should be null. Thus, in this study we observed apparent "group polarization" (see, e.g., Isenberg 1986) wherein like-minded discussion groups move further in the direction of their initial (partisan) tendency. This might have been attributable to the excess loquaciousness of extremists observed in our test of Hypothesis 2, so we conducted additional analyses (see Table 9) of endline affective polarization in which participants' partners' baseline attitudes are interacted with binary indicators for whether the partner was more loquacious than the participant themself. The variable "Partner Sent More Messages" is 1 when the participant's partner sent more messages than the participant themself sent (and 0 otherwise), and the variable "Partner Wrote More Characters" is 1 when the participant's partner wrote more characters than the participant themself wrote (and 0 otherwise). In neither operationalization does loquaciousness appear to moderate persuasive influence. Rather, as shown in Model 3 of Table 10, it appears that the co-partisan accountability treatment may have been responsible for some polarization of participants' partisan affect (indicated by the negative coefficient on the neutral accountability condition).

13. To infer causality in a counterfactual sense, one would need an experiment with a no-chat control.
14. Given random group composition, baseline affect can be considered a randomized treatment on one's partner.

Table 9. Group Polarization (Interaction Tests of Relative Loquaciousness)

	Dependent variable:	
	Post-Chat Partisan Affect (D - R Feeling Thermo Diff)	
	(1)	(2)
Pre-Chat Affect	0.959*** (0.017)	0.961*** (0.017)
Partner's Pre-Chat Affect	0.077** (0.033)	0.083*** (0.030)
Partner Sent More Messages	1.626 (2.261)	
Partner's Pre-Chat Affect × Partner Sent More Messages	0.009 (0.038)	
Partner Wrote More Characters		–0.555 (2.376)
Partner's Pre-Chat Affect × Partner Wrote More Characters		–0.005 (0.040)
Constant	–0.025 (2.079)	0.891 (1.608)
Observations	362	362
R^2	0.880	0.878
Adjusted R^2	0.878	0.877
Residual Std. Error (df = 357)	9.951	9.999

Note: *p<0.1; **p<0.05; ***p<0.01

Table 10. Group Polarization

	Dependent variable:		
	Post-Chat Partisan Affect (D - R Feeling Thermo Diff)		
	(1)	(2)	(3)
Pre-Chat Affect	0.969*** (0.016)	0.961*** (0.017)	0.962*** (0.017)
Partner's Pre-Chat Affect		0.080*** (0.021)	0.082*** (0.021)
Treatment (Neutral Accountability)			–1.829* (1.014)
Constant	4.615*** (1.069)	0.613 (1.329)	1.318 (1.495)
Observations	389	362	362
R^2	0.881	0.878	0.879
Adjusted R^2	0.880	0.878	0.878
Residual Std. Error	9.921 (df = 387)	9.981 (df = 359)	9.953 (df = 358)

Note: *p<0.1; **p<0.05; ***p<0.01

Table 11. Group Polarization (Interaction Tests with Accountability Treatment)

	Dependent variable:	
	Post-Chat Partisan Affect (D - R Feeling Thermo Diff)	
	(1)	(2)
Pre-Chat Affect	0.963*** (0.017)	0.974*** (0.017)
Treatment (Neutral Accountability)	2.739 (2.238)	–6.858 (4.545)
Partner's Pre-Chat Affect	0.117*** (0.031)	
Partner's Pre-Chat Affect × Treatment	–0.081** (0.040)	
Partisanship Strength		–1.344 (1.487)
Partisanship Strength × Treatment		2.044 (1.793)
Constant	–0.706 (1.855)	8.481** (3.651)
Observations	362	389
R^2	0.881	0.882
Adjusted R^2	0.879	0.881
Residual Std. Error	9.905 (df = 357)	9.910 (df = 384)
Note:		*p<0.1; **p<0.05; ***p<0.01

Table 11 presents additional supplementary models in which the treatment indicator is interacted with the participant's partner's pre-chat affect, and with the participant's own pre-chat partisanship strength, respectively. The interaction of the neutral accountability treatment with partner's pre-chat affect was found to be statistically significant and negative, which suggests that the co-partisan treatment might have caused study participants with particularly strong partisan affect to express themselves more stridently, polarizing their partner in the process (although we must interpret this analysis with caution, since it was not pre-registered and concerns an interaction effect). No evidence was found of an interactive relationship between the treatment and participants' pre-chat strength of partisanship.

As an exploratory analysis of polarization mechanisms, we analyzed the sentiment of each message (by wrapping `sentiment()` from the `sentimentr` package within `rechat`'s `featurizeChat()` function), and constructed summary measures of each participants' overall sentiment by taking the mean sentiment score over all of each participant's messages (with `rechat`'s `summarizeChat()` function). We then analyzed predictors of message sentiment in a linear regression framework (with

standard errors clustered at the chatroom level), and found that baseline partisan affect was a strongly significant predictor of mean message sentiment in the chat. This is consistent with a persuasion mechanism for changes in partisan affect over the course of the chats. We did not find that one's message sentiment was predictive of one's partner's post-chat partisan affect, however. Future studies should probe more deeply how chat content mediates processes of attitude change.

Table 12. Baseline Partisan Affect and Mean Message Sentiment

	Dependent variable:	
	Mean Sentiment of Messages	
	(1)	(2)
Baseline Partisan Affect	–0.001*** (0.0002)	–0.001*** (0.0003)
Extroversion		0.003 (0.006)
Self-Monitoring		0.008 (0.006)
Political Interest		–0.005 (0.006)
6-Point PID		–0.020 (0.015)
7-Point Ideology		–0.005 (0.008)
Ideological Identity		0.003 (0.009)
Media Consumption Scale		0.002 (0.007)
Constant	0.025* (0.013)	0.096** (0.048)
Observations	362	362
R^2	0.033	0.052
Adjusted R^2	0.031	0.031
Residual Std. Error	0.111 (df = 360)	0.111 (df = 353)
Note:		*p<0.1; **p<0.05; ***p<0.01

Appendix 11.2 Discussion

Study 2 offers a sample of the kinds of analyses afforded by chat-based research. Our finding that ideologues were more loquacious than moderates is highly pertinent to political polarization, and our other analyses raise a variety of interesting questions for future research.

Ideologues' loquaciousness is thought to polarize online political discourse: although most people hold moderate views, extremists self-select into expressing their views at higher rates than moderates, thus distorting the sample of opinions that are presented online (e.g. Bor and Petersen 2021; Hughes 2019; Bail 2021). Extremists' loquaciousness may also contribute to partisans' tendency (e.g. Druckman et al. 2022) to overestimate each others' extremity. The fact that this pattern replicates in the

ReChat environment indicates it is quite robust, and future research should probe its causes further.

Extremists' loquaciousness also might contribute to group polarization – the tendency for discussions to lead like-minded groups to become more extreme (e.g. Myers and Lamm 1976, see also Klar 2014 and Druckman, Levendusky, and McLain 2018) – which we also observed in this study. Burnstein and Vinokur (1977) theorized that group polarization occurs because extremists make more arguments than moderates, consistent with our observation that extremists were more loquacious. However, while we *did* find evidence of persuasion, we did *not* see evidence that persuasion was moderated by loquaciousness. Instead, we found suggestive evidence that the co-partisan accountability treatment may have contributed to polarization, which is more consistent with Turner's (1991) theory of group polarization as an effect of group prototypes: anticipating judgment by fellow Democrats may have polarized chats and subsequent affect through a normative mechanism (see also Abrams et al. 1990). This represents an intriguing subject for future chat studies.

Equally intriguing is the fact that our pre-registered analysis of *character* count found null results for our gender and accountability-treatment hypotheses and rejected the null for our ideology hypothesis, while exploratory analyses of *message* count gave inverse results. We tentatively propose that message count and character count reflect substantively distinct behaviors: writing a large number of characters arguably reflects having a large number of considerations that come to mind (Zaller 1992), and feeling comfortable expressing them in a given social context (Noelle-Neumann 1991). We speculate that liberal Democrats are likely to have more opinions *that they are comfortable expressing* among fellow Democrats, while moderates may be more prone to self-censorship, with the result that they have less to say overall.

Sending many messages, meanwhile, may be an expression of social dominance: it may afford the sender greater control over the topic and framing of a discussion, independent of how many ideas they actually have. This is consistent with a social dominance interpretation of gender effects in deliberation (Karpowitz and Mendelberg 2014), and their manifestation in the specific behavior of negative interruptions (Mendelberg, Karpowitz, and Oliphant 2014). It is also consistent with the logic that the accountability treatment cannot plausibly give participants more ideas to express about the topic, but *could* affect participants' desire to "seem thorough," which they may achieve by sending more messages. To the extent that the thoroughness incentive implicitly puts participants in competition with each other, it is also plausible that variations in this incentive might induce

variations in dominance displays, which would be a good subject for additional research.

Appendix 12. Study 2 Outcome Skew and Robustness Tests

Count data often exhibits right-tailed skewness, with a concentration of low values and a long tail of high values. This can lead to non-normality of residuals and potentially biased estimates in linear regression analyses. This appendix investigates this possiblity and conducts additional analyses as a robustness check.

Figure 10 plots histograms of the four outcomes considered in our main analyses: character count (the pre-registered outcome), message count, word count, and unique word count. A D'agostino (1970) test found significant skewness in all four outcomes, but as is visible in Figure 10, message count was particularly skewed.

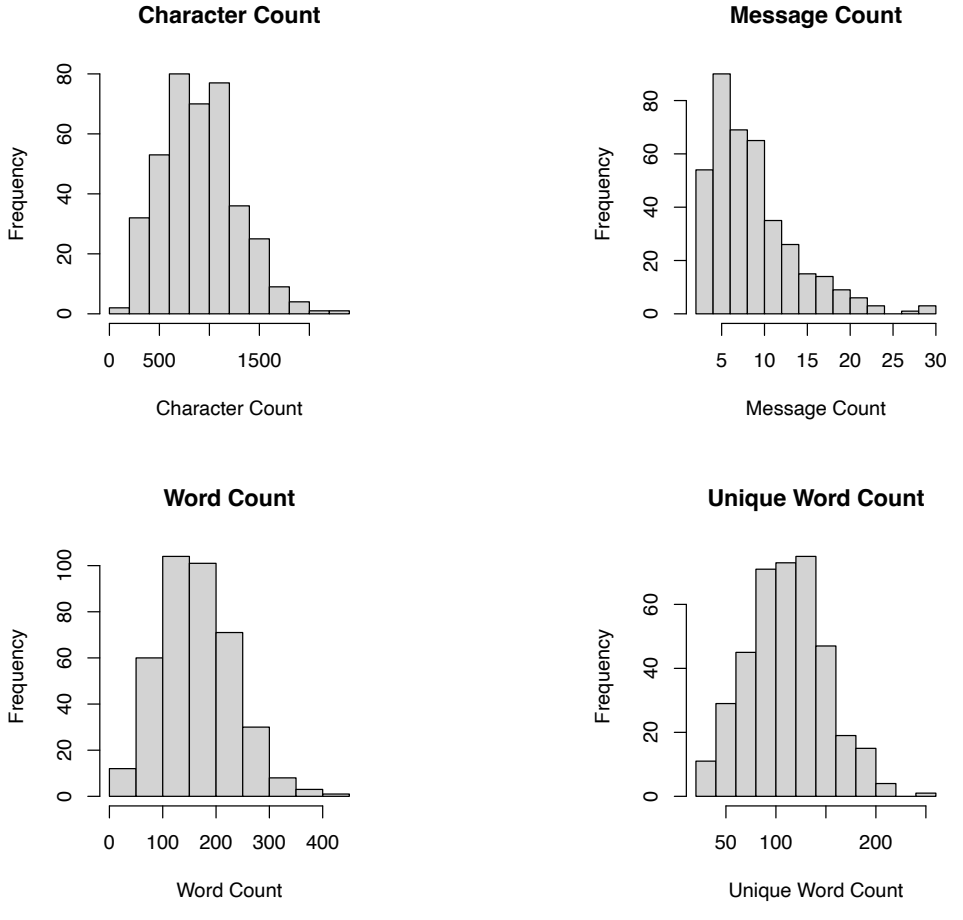


Figure 10. Study 2 outcome histograms showing skew.

We therefore implemented log-linear versions of our models, as a robustness check on the loquaciousness analyses presented in the main text. We transformed our count outcomes using a natural logarithm and re-estimated our models with the original predictors. As shown in Table 13, these analyses produced essentially the same results as our main analyses (presented in Table 8). This consistency reassures us regarding the robustness of our results.

Table 13. Logged Loquaciousness (Study 2)

	Dependent variable:			
	Log # Characters (Pre-Reg)	Log # Messages	Log # Words	Log # Unique Words
	(1)	(2)	(3)	(4)
Treatment (Neutral Accountability)	0.033 (0.047)	0.148*** (0.057)	0.047 (0.047)	0.026 (0.040)
7-Point Ideology	–0.068** (0.030)	–0.057* (0.031)	–0.065** (0.029)	–0.061** (0.024)
Male	–0.034 (0.051)	0.122** (0.055)	–0.039 (0.051)	–0.022 (0.042)
Constant	6.838*** (0.066)	2.063*** (0.078)	5.130*** (0.066)	4.767*** (0.054)
Observations	390	390	390	390
R^2	0.016	0.036	0.016	0.017
Adjusted R^2	0.008	0.028	0.009	0.009
Residual Std. Error (df = 386)	0.477	0.530	0.476	0.396
F Statistic (df = 3; 386)	2.029	4.768***	2.119*	2.238*

Note: *p<0.1; **p<0.05; ***p<0.01

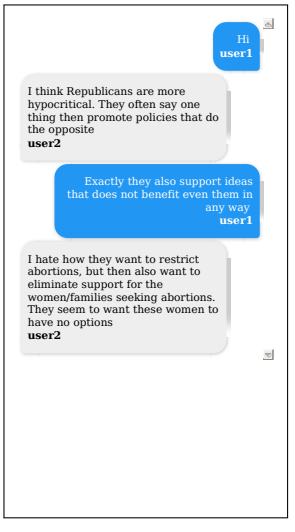
Appendix 13. Example Chats

This appendix displays 12 chats: 6 chats randomly-selected from Study 1, and 6 randomly-selected from Study 2. These examples are intended to illustrate the quality, length, and structure of the interactions studied in this paper. PDF exports of all chats are available in the online supplementary materials.

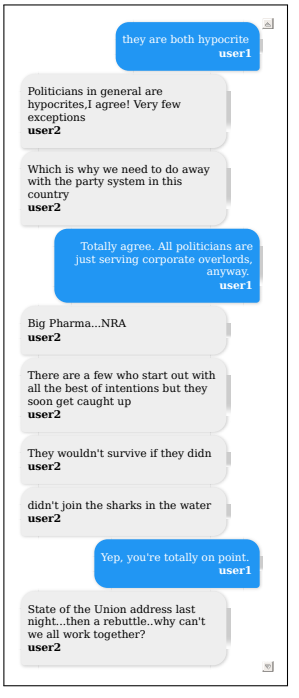
Appendix 13.1 Study 1 (5 Minute Chats)

Figures 11 and 12 display examples of chats conducted in Study 1, where the chat time limit was set to 5 minutes. Although both participants were able to express their views in all these examples, only Examples 2 and 5 feature a repeated "back and forth" conversation between the two participants, indicating that 5 minutes may be shorter than ideal for generating true conversations.

(a) Example 1



(b) Example 2



(c) Example 3

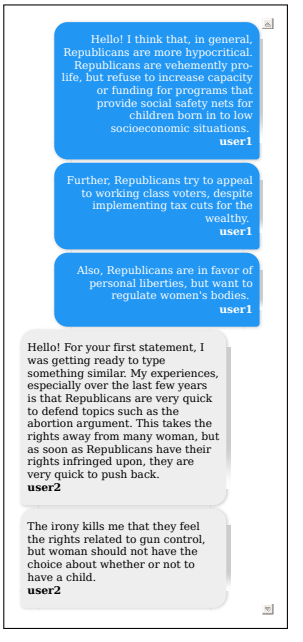
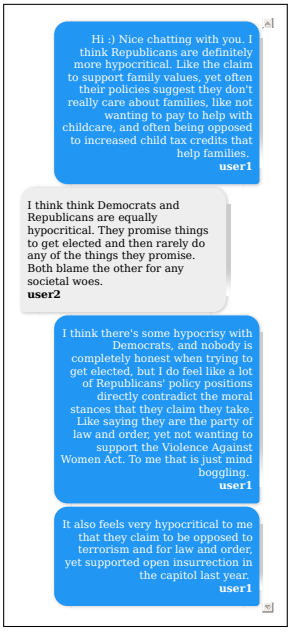
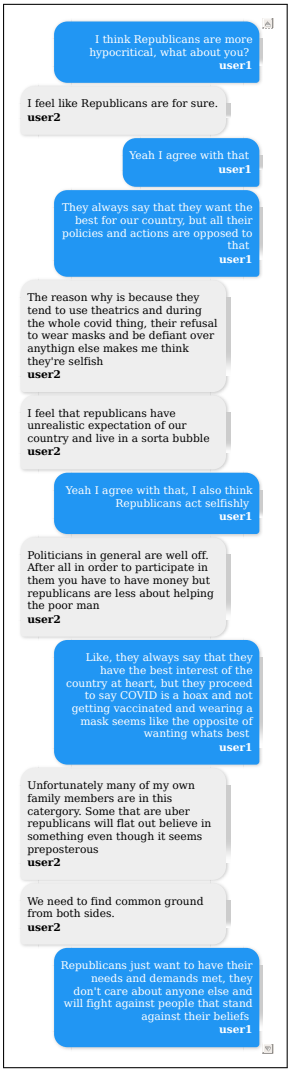


Figure 11. Example Chats from Study 1

(a) Example 4



(b) Example 5



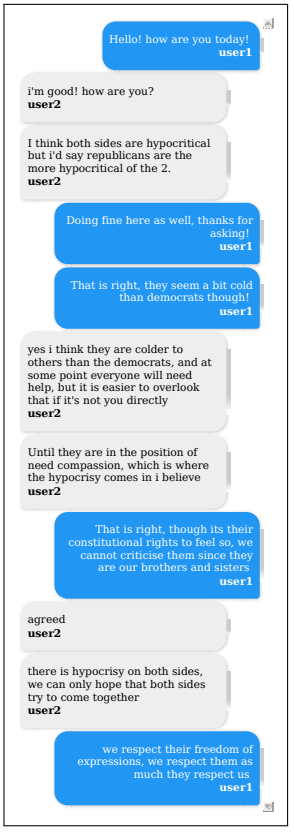
(c) Example 6

Figure 12. Example Chats from Study 1 (Continued)

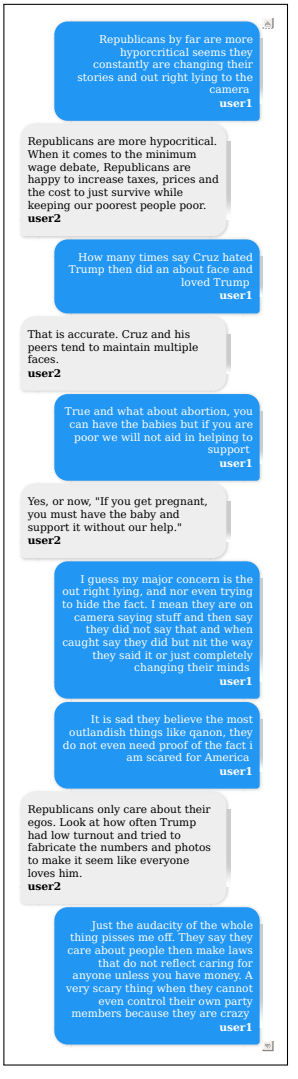
Appendix 13.2 Study 2 (10 Minute Chats)

Figures 13 and 14 display examples of chats conducted in Study 2, where the chat time limit was set to 10 minutes. All of these examples feature a robust conversational exchange between the participants, suggesting that 10 minutes may be a more suitable duration than 5 minutes, for collecting true "conversations."

(a) Example 1



(b) Example 2



(c) Example 3

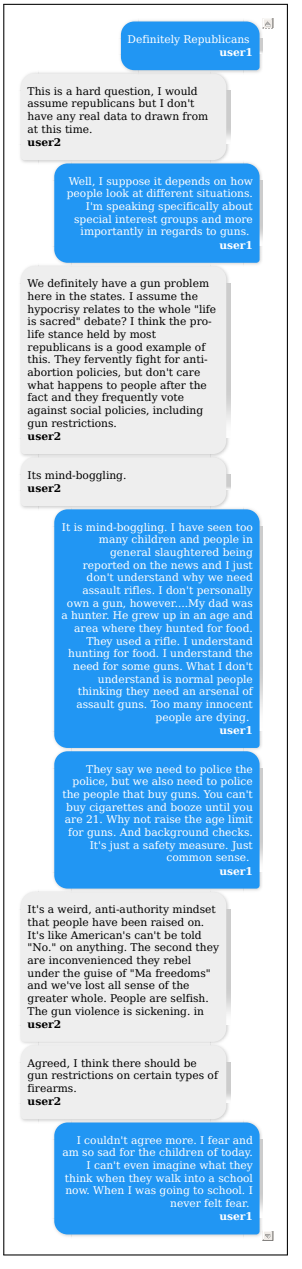


Figure 13. Example Chats from Study 2

(a) Example 4

I think Republicans are more hypocritical. All their talk of "Pro-Life" yet they make no laws or programs to support mothers or children after birth.
user1

I think both parties fall victim to being hypocritical. This is evident through daily news stories. One party will point a finger at the other for wrongdoing and then they will go and do something similar the next day.
user2

We need universal maternity leave, free infant care, free pre-k, the list goes on.
user1

I do agree with your point too Kimber.
user2

There is so little that is done to help mothers and mothers to be.
user2

And the recent school shooting in Texas highlights how many Republicans refuse to do anything about gun control
user1

All of their policies seem to be about control of others' lives but do nothing to support the people in their daily lives or struggles.
user1

That is also a very good point. It is a devastating problem with too weak of answers for us right now. As a mother of three. It is absolutely terrifying.
user2

They want to try to shift the blame to 'mental health' or arming teachers- which is a ridiculous idea- instead of enforcing age limits on gun purchases, etc.
user1

I'm a mother myself and I'm scared to send my little one to school.
user1

We're the only country in the world that had this issue.
user1

More than age limits need to be enforced. 18 years old is certainly not a developed brain. It is sad.
user2

I agree. Do you think Republican voters agree with their representatives' policies?
user1

I do not. I feel like many fof us are actually much more aligned in the middle.
user2

In my state at least, it seems as though they vote against their own self interests.
user1

Yes, I see that as well. There is alot of wanting to "win" rather than wanting to truly fix things.
user2

I would have thought that most people were in the middle some years ago but the Republican party has seemed to swing far far right these days
user1

I agree with you on the wanting to win aspect. I wish bipartisanship was a reality in more topics.
user1

(b) Example 5

Republicans
user1

Hello! I would agree with that
user2

Yessir, why would you say?
user1

They seem to always push things for "family values" regardless of how it affects real american families accross the nation
user2

I agree 100%.
user1

Is there anything specific you would like to add?
user2

I also feel like their hypocrisy towards the COVID 19 vaccines saying they should be able to do what they would like with their bodies is pretty hypocritical in conjunction with their views on abortions
user2

Itâ€™s just that, and they donâ€™t really care about as much as they try to make it seem. Like the abortions, nobody should have the right to tell someone what they can and canâ€™t do wit their bodies so i would have to agree on that too!
user1

(c) Example 6

democrats care more about feelings now while republicans want to keep things the same as the country was
user1

so i say republicans are more hypocritical
user1

Both can be hypocritical for sure, but we are seeing it right now with republicans who say no to gun control because shootings will just happen anyway, but then they support abortion even though abortions will happen
user2

especially when they tell ppl they cant have an abortion but complain when told to put on a mask saying its their body
user1

Yes, I was just thinking of that also
user2

Or belief in specific Biblical principles but helping the poor or serving others isn't happening too much
user2

Or let's accuse teachers of brainwashing and try to ban books and control the curriculum because they don't trust them but then they also say that teachers need to be armed. I am a teacher, and the day that happens, teachers will walk out in droves
user2

I feel like biblical principles are always hypocritical, my girlfriends mom is really religions and preaches about love and peace but is one of the rudest people i know
user1

Yep, ask any server who tips the worst and it's the after-church crowd on Sundays
user2

gun regulation can just be left in the hands of those who are trained to use them properly, and if bad people still get the guns then its the "good" people giving it to them
user1

but also people who love the 2nd amendment want the guns to fight against the country they love so much if the country wants to take them away
user1

I am seeing a lot today about how the police responders failed to really do their jobs well. I don't want to be an armchair quarterback, but if they are afraid to go head to head against someone with those kind of weapons, how can anyone else?
user2

Yes, that's some hypocritical patriotism for sure
user2

police choose to be in that position, to be fought and hated but also respected. when people say blue lives matter its dumb since they can choose to take off the blue, black lives cannot
user1

if they can't handle the heat or backlash to do their job properly than its time to quit
user1

Figure 14. Example Chats from Study 2 (Continued)

Appendix 14. Human Subjects Research

This research entailed minimum risk to human subjects. The only arguable "deception" – the manipulation of the identity of the thoroughness judges as "Democratic" or "Neutral" – did not undermine the safety or welfare of subjects. Therefore no debrief was conducted. Subjects gave informed consent for participating in a survey and a chat, separately. Participants were paid $1 for participating in the survey, and were paid a $1 bonus for participating in the 5-minute chat in Study 1, and a $2 bonus for participating in the 10-minute chat in Study 2. Participants whose chats were above median thoroughness were granted an additional $1 bonus. This compensation was chosen to exceed the federal minimum wage for the amount of time participants were expected to spend on the studies. Participants were asked to provide pseudonymous nicknames to use during the chat, to mitigate potential threats to privacy.

References

Abrams, Dominic, Margaret Wetherell, Sandra Cochrane, Michael A. Hogg, and John C. Turner. 1990. Knowing what to think by knowing who you are: Self-categorization and the nature of norm formation, conformity and group polarization*. *British Journal of Social Psychology* 29 (2): 97–119.

Bail, Christopher A. 2021. *Breaking the social media prism: how to make our platforms less polarizing.* 1st. Princeton: Princeton University Press.

Bor, Alexander, and Michael Bang Petersen. 2021. The Psychology of Online Political Hostility: A Comprehensive, Cross-National Test of the Mismatch Hypothesis. *American Political Science Review,* 1–18.

Broockman, David E, Joshua L Kalla, and Sean J Westwood. 2023. Does Affective Polarization Undermine Democratic Norms or Accountability? Maybe Not.

Burnstein, Eugene, and Amiram Vinokur. 1977. Persuasive argumentation and social comparison as determinants of attitude polarization. *Journal of Experimental Social Psychology* 13 (4): 315–332.

D'Agostino, R.B. 1970. Transformation to Normality of the Null Distribution of G1. *Biometrika* 57 (3): 679–681.

Druckman, James N., Samara Klar, Yanna Krupnikov, Matthew Levendusky, and John Barry Ryan. 2022. (Mis-)Estimating Affective Polarization. *The Journal of Politics,* 1106–1117.

Druckman, James N., Matthew S. Levendusky, and Audrey McLain. 2018. No Need to Watch: How the Effects of Partisan Media Can Spread via Interpersonal Discussions. *American Journal of Political Science* 62 (1): 99–112.

Groenendyk, Eric, and Yanna Krupnikov. 2021. What Motivates Reasoning? A Theory of Goal-Dependent Political Evaluation. *American Journal of Political Science* 65 (1): 180–196.

Hughes, Adam. 2019. *A small group of prolific users account for a majority of political tweets sent by U.S. adults.*

Isenberg, D. J. 1986. Group polarization: A critical review and meta-analysis. *Journal of Personality and Social Psychology* 50 (6): 1141–1151.

Karpowitz, Christopher, and Tali Mendelberg. 2014. *The Silent Sex: Gender, Deliberation, and Institutions.* Princeton: Princeton University Press.

Klar, Samara. 2014. Partisanship in a social setting. *American Journal of Political Science* 58 (3): 687–704.

Lee, David S. 2009. Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects. *REVIEW OF ECONOMIC STUDIES.*

Marwick, Alice E., and Danah boyd. 2011. I tweet honestly, I tweet passionately: Twitter users, context collapse, and the imagined audience. *New Media & Society* 13 (1): 114–133.

Mendelberg, Tali, Christopher F. Karpowitz, and J. Baxter Oliphant. 2014. Gender inequality in deliberation: Unpacking the black box of interaction. *Perspectives on Politics* 12 (1): 18–44.

Myers, David G., and Helmut Lamm. 1976. The group polarization phenomenon. *Psychological Bulletin* 83 (4): 602–627.

Noelle-Neumann, Elisabeth. 1991. The Theory of Public Opinion: The Concept of the Spiral of Silence. In *Communication Yearbook,* 256–287. International Communication Association.

Roberts, Margaret E., Brandon M. Stewart, Dustin Tingley, Christopher Lucas, Jetson Leder-Luis, Shana Kushner Gadarian, Bethany Albertson, and David G. Rand. 2014. Structural Topic Models for Open-Ended Survey Responses: STRUCTURAL TOPIC MODELS FOR SURVEY RESPONSES. *American Journal of Political Science* 58 (4): 1064–1082.

Santoro, Erik, and David E. Broockman. 2022. The promise and pitfalls of cross-partisan conversations for reducing affective polarization: Evidence from randomized experiments. *Science Advances* 8 (25): eabn5515.

Tauchmann, Harald. 2014. Lee (2009) Treatment-Effect Bounds for Nonrandom Sample Selection. *The Stata Journal: Promoting communications on statistics and Stata* 14 (4): 884–894.

Turner, John C. 1991. *Social influence.* Bristol, PA: Open University Press.

Zaller, John. 1992. *The Nature and Origins of Mass Opinion.* Cambridge: Cambridge University Press.