Fora: A corpus and framework for the study of facilitated dialogue

Hope Schroeder[†‡], **Deb Roy**[†‡], **Jad Kabbara**[†‡]
[†]Massachusetts Institute of Technology, [‡]MIT Center for Constructive Communication
{hopes, dkroy, jkabbara}@mit.edu

Abstract

Facilitated dialogue is increasingly popular as a method of civic engagement and as a method for gathering social insight, but resources for its study are scant. We present Fora, a unique collection of annotated facilitated dialogues. We compile 262 facilitated conversations that were hosted with partner organizations seeking to engage their members and surface insights regarding issues like education, elections, and public health, primarily through the sharing of personal experience. Alongside this corpus of 39,911 speaker turns, we present a framework for the analysis of facilitated dialogue. We taxonomize key personal sharing behaviors and facilitation strategies in the corpus, annotate a 25% sample (10,000+ speaker turns) of the data accordingly, and evaluate and establish baselines on a number of tasks essential to the identification of these phenomena in dialogue. We describe the data, and relate facilitator behavior to turn-taking and participant sharing. We outline how this research can inform future work in understanding and improving facilitated dialogue, parsing spoken conversation, and improving the behavior of dialogue agents.

1 Introduction

In light of rampant toxic polarization on social media (Bail, 2022), sliding trust in democratic institutions, and the shortcomings of traditional methods of seeking and using public opinion, new methods for responsibly seeking participatory insight are critically needed across many fields, including for civic use in governance and for nuanced insight in the social sciences.

Social media has been used as a source of insight into public experience and discourse by the natural language processing (NLP) and computational social science communities for decades now, but well-documented user self-selection and polarization make trusting insights drawn from social media difficult. While surveys and other standard social research methods have strong empirical grounding, they often limit opportunities for free-form comment by participants.

Facilitated dialogue provides one alternative avenue of seeking input and insight from various stakeholders, community members, constituents, or other participants. One example use case for facilitated dialogue is for community listening. Members of the community can be invited to participate in small group conversations, where they are prompted by a trained facilitator to share their personal experiences and perspectives regarding a specific topic that is important to the community. Analysis of patterns across conversations can serve as useful input to decision-making processes that impact the community, such as public policy formation or a hiring process for a public servant. Participation in such community engagement processes may also strengthen "civic muscle."

In this work, we present Fora,[1] a corpus of annotated transcripts of 262 multi-person facilitated dialogues similar to the one outlined above. These conversations were organized with various US-based organizations that partnered with the nonprofit Cortico. Their goal was to engage members or stakeholders by inviting them to converse in an effort to understand their personal experiences and perspectives on topics like education, public health, and upcoming elections. Along with this corpus of facilitated dialogue, we present a general framework for the analysis of facilitated group dialogue through outcome metrics relevant to evaluating a facilitation process, including turn-taking. We taxonomize key sharing behaviors and facilitation strategies. We run a human annotation effort for a 25% (10,000+ speaker turns) subset of the corpus. We use it to train or prompt and evaluate the performance of several language models automating the identification of facilitation strategies

[1]Accessible at https://github.com/schropes/fora-corpus

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 13985–14001

and sharing behaviors, thus establishing baselines on tasks essential to the study of facilitated dialogue. Moreover, we describe the data in terms of facilitator behavior and relate facilitator behavior to turn-taking and participant sharing. We outline how this research can inform work to understand and improve facilitated dialogue for facilitators and participants, and we discuss its usefulness in future work on parsing spoken conversations, and improving dialogue agents.

2 Related Work

The study of conversation spans many disciplines including linguistics, NLP, and the social sciences. Building on speech act theory, major efforts to taxonomize, annotate, classify, and model dialogue acts have been important to the foundation of NLP. Renowned contributions include the Switchboard Corpus (Godfrey et al., 1992) taxonomy of speech acts, DAMSL corpus (Core and Allen, 1997), early work from Stolcke et al. (2000) on tagging speech acts in dialogue, and DialogBank, an interoperable system of dialogue act annotation (Bunt et al., 2019).

In parallel, scholars in other fields have noted the importance of dialogue to public understanding, especially in the face of social conflict. Bohm and Nichol (2004) lay the groundwork for academic work on understanding both the value of facilitated dialogue for collective understanding, as well as the unique role of the facilitator in setting the norms and guiding the conversation towards emergent expression. The field of deliberative polling makes a case for the importance of deliberative discourse in leading to better decision outcomes (Fishkin, 1997) and the legitimization of democratic outcomes (Matravers and Pike, 2005). Saunders (1999) discusses the importance of sustained dialogue for changing human relationships in the face of deeply-rooted conflict, and Ryfe (2006) notes the importance of storytelling in productive deliberation forums. Others note the power of narrative for digital participation (Esau, 2018) and democracy (Black, 2013).

Corpora of public dialogue outside the social media space are rare, and high-quality annotated spoken conversation transcripts are even rarer. Word distribution and turn-taking patterns in speech and transcribed speech are distinct from written text, making spoken speech datasets for understanding community insight important. The recent launch of the CANDOR Corpus (Reece et al., 2023) marks an important leap in conversation research. EuroPolis (Gerber et al., 2018), a set of discussions from a transnational European deliberative poll, provides another related dataset, and research from Dillard (2013) finds a relationship between facilitation types and the trajectory of public deliberations.

Dialogue can function as a method, tool, or input to a process of gradual understanding of a problem, including in a research process. Many fields of social research use similar methods of surfacing insight in a group setting, including group interviews, discussion groups, and focus groups. Focus groups are a common and validated method of surfacing insight from a group (Stewart and Shamdasani, 2014) and they have existed for 80 years (Onwuegbuzie et al., 2009). There are documented and commonly used practices for running focus groups (Onwuegbuzie et al., 2009) but to our knowledge, there are no known datasets for their study at scale. Corpora of recorded meetings like one described in Bates et al. (2005) exist, but facilitated conversations differ from meetings in important ways.

A rich theory on facilitation techniques comes from literature in education. Research from Motozawa et al. (2021) shows that facilitative moves can have an impact on the outcome of interest to the facilitation setting, and that "request" utterances by the facilitator encourage response. Research from Chung (2011) on education discussion fora relates the presence of nine facilitation techniques on those fora, including summarization, making connections, providing opinions, and inviting feedback, to the presence of critical thinking by participants in online discussions. Research on facilitation for teaching describes three main functions of facilitation: social, organizational, and intellectual (Cheung and Hew, 2010). The function role of the facilitator, according to Kolb et al. (2008) is to manage group discussions and processes so that group members have a positive experience, to host a process that "promotes valuable results in group dialogue, analysis, and planning" and to use techniques that aid in group interaction or the accomplishment of those goals.

3 Motivation

Hand-transcribed multi-person conversation datasets of substantial size are relatively rare in the NLP community, and to our knowledge, no corpora of annotated and analyzed semi-structured facilitated dialogues currently exist. Furthermore, as

online discussion platforms look to promote civil discourse, experiments in using automated agents to facilitate and moderate discussion or conduct interviews are already underway (Xiao et al., 2020; Kim et al., 2021; Shiota et al., 2018; Schluger et al., 2022). As such, understanding human facilitation is crucial to the successful development of productive conversations, including ones that may involve an automated agent. As facilitated dialogue emerges as a forum for discourse and public input in contrast to discourse online, a better understanding of facilitation behaviors can inform the design of productive group conversations, no matter their intended outcome or purpose.

4 Corpus

The Fora corpus is a set of conversation transcripts, conversation guides, and annotations associated with 262 conversations spread across 16 conversation collections convened over a variety of topics in diverse communities. Each conversation collection was convened by a partner organization in the United States which, in partnership with Cortico, a US-based social technology non-profit, hosted a series of conversations of similar structure about a community issue or series of issues. Conversation collections range in size from a minimum of 5 conversations to a maximum of 76 conversations. Each conversation collection has a companion conversation guide, developed by the hosting partner organization in partnership with Cortico. Facilitator roles in the conversation were hand-labeled by our researchers. Most conversations are facilitated by one individual. 29 are co-facilitated, of which 28 of which were facilitated by 2 individuals and a single one was co-facilitated by three individuals.

Conversations in the Fora corpus occurred between November 14th, 2019, and April 6th, 2023. Conversations were held with a median of 6 people, with a minimum of 2 and maximum of 12. The average conversation length was 66 minutes, with a minimum of 8 minutes and a maximum of 117 minutes, or just under 2 hours. Conversations had an average of 152 speaker turns, with a minimum of 14 turns and a maximum of 719. Most conversations were held online over meeting software. Conversations were recorded with consent and transcribed by REV transcription software, and partners were able to redact and edit transcripts for mistakes following the conversation. A visualization of the duration of speaker turns across participants in a typical conversation in the Fora corpus can be seen in Figure 1. The appendix provides more descriptive statistics on the conversation collections within the corpus.

4.1 Conversation structure

Each conversation collection in the corpus is matched to a conversation guide that facilitators were instructed to use for the conversation. Conversation guides were developed in partnership with the partner organization involved in hosting the conversations. Conversation guides follow a similar structure across conversation collections, adapted to the subject area of interest to the partner organization or the purpose of the hosted conversations.

Questions in the conversation guide are intentionally open-ended and are designed to encourage participants to share personal stories and personal experiences. In contrast to top-down approaches commonly seen in focus groups, this semi-structured approach is intended to lead to an agenda that emerges organically through personal sharing. Facilitators often begin by setting context through a script, and read a consent statement before stating conversation norms. The facilitator then transitions to a request for sharing. For example, one first prompt from a conversation guide is: "Take a minute and think of a memory or a personal story from your life that has shaped who you are, and would help others understand what is important to you." Then, in response, participants transition to community-focused conversation. The conversations often end with a facilitator prompt like "What is one change you would most like to see happen?" Participants are then encouraged to ask any questions before ending the session.

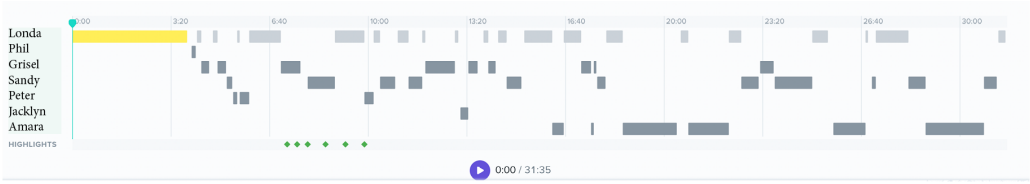
5 Characterizing Facilitated Dialogue

We first characterize facilitated dialogue using a number of computational metrics to understand variation in the corpus across collections.

5.1 Inequality in time-sharing

Turn-taking is a natural part of human conversation and modeling it is central to the study of conversation (Wilson et al., 1984). Coordination behaviors in conversation can vary across cultures (Stivers et al., 2009) and genders (Ghilzai, 2015). In the context of facilitated dialogue, the facilitator has an active role to play in setting, encouraging, and enacting norms for the conversation. Because

Figure 1: Example visualization of conversation audio in a Fora conversation. In it, "Londa" is the facilitator, starting off the conversation before participants take turns through the conversation.



the facilitated dialogues in this dataset exist as a method of providing space to share insight, statistics describing how limited time is shared across conversation participants is a useful heuristic for characterizing the nature of the conversation. We operationalize equitable distribution of time in the facilitated conversation, a limited resource, with a classic metric of inequality over limited resources from the economics literature, the Gini coefficient, as follows:

$$1 - G = 1 - \frac{1}{n(n-1)\bar{x}} \sum_{i=1}^{n} (n + 1 - 2i)x_i \quad (1)$$

where n is the number of observations (participants in conversation), i is the index of the observation in array x, x_i is the resource available to the ith person, and $\bar{x}$ is the mean of the array x containing all participants and their resources.

For each conversation, we model the equitable sharing of time in the conversation using the Gini coefficient in two ways. First, we treat the total summed duration of a participant's speaking time across all conversation turns as the share x_d of our variable of interest, time. Next, we treat the total count of a participant's speaking turns as the share x_c of our variable of interest, time. For conversation C, we use the equation to give us a representation G_d of the Gini coefficient with respect to the sharing of time duration in conversation C and a representation G_c of the Gini coefficient with respect to the sharing of turn count in conversation C. We plot G_d and G_c to visualize the spread of time and turn count inequality across conversations. A higher value for the Gini coefficient indicates higher inequality in resource sharing within the conversation.

We see in Figure 2 there is wide spread in the Gini coefficients of time-sharing among participants across conversations and collections, indicating variation in the distribution of time across participants, both in terms of turn count, and in duration of speaking time.

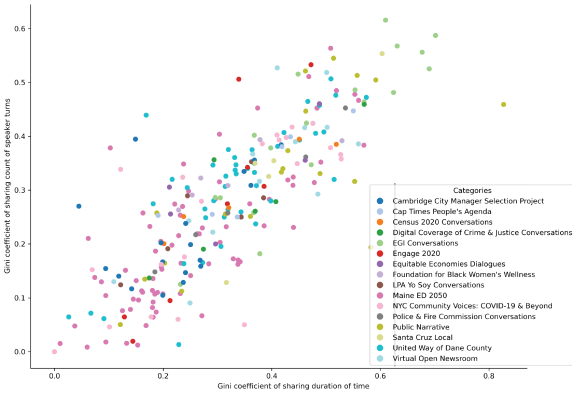


Figure 2: Time-sharing in Fora conversations across collections

There are many possible explanations for this variation: some conversation participants may be more interested in sharing than others, and some participants may take longer to share a single story than their more parsimonious counterparts.

Facilitators naturally respond to the conditions of the conversation, including to the personalities of participants, in order to accomplish their goals for the conversation. As such, while there are shortcomings to any measurement of time-sharing across participants, understanding the spread of time-sharing within the collection is useful. Its relationship to participant sharing will be discussed later.

5.2 Participant & facilitator turn-taking ratios

Another useful descriptive measure of turn-taking describes the amount of conversation taken up by the facilitator compared to the participants. Again, we can operationalize a metric describing the ratio of facilitator speaking duration to average participant speaking duration within a conversation.

The first ratio R_d is defined as:

$$R_d = \frac{F_d}{\bar{P}_d},$$

where:

R_d is the ratio of facilitator speaking duration to average participant speaking duration,
F_d is the facilitator speaking duration,
$\bar{P}_d$ is the average participant speaking duration.

The second ratio R_c is defined as:

$$R_c = \frac{F_c}{\bar{P}_c},$$

where:

R_c is the ratio of facilitator speaking turn count to average participant speaking turn count,
F_c is the count of facilitator speaking turns,
$\bar{P}_c$ is the average count of participant speaking turns.

We can visualize the spread of and relationship between these measures by plotting R_d and R_c, as seen in Figure 4 in the appendix to give another high level view of time-sharing in the corpus.

5.3 Average participant turn length

Styles of facilitated dialogue differ. Some facilitated dialogues are a series of long participant turns with minimal facilitator intervention to hand off between participant monologues. Others appear more like a casual back-and-forth conversation. We calculate average participant turn length by taking the subset of speaker turns spoken by participants. Doing so, we find the mean average turn length across collections is 41 seconds. The lowest average turn length in a conversation is 4.7 seconds, and the highest average turn length is 130 seconds, or over two minutes. These two extremes exemplify different conversation styles: the participants in the one with the lowest average turn length gave quick answers, whereas the participants in the longest one told unbroken stories in a series of monologues. As such, a high-level statistic for average turn length can help characterize conversation style. A figure visualizing collections by average turn length can be found in Figure 5 in the appendix. We use a one-way ANOVA to show a statistically significant difference across collections regarding average participant turn length ($p < .01$). Factors affecting the average turn length could include the level of comfort of participants with each other (perhaps there is more casual interruption among friends) or the preferred style and norm-setting done by the facilitator (some facilitators manage conversations as a sequence of monologues). Follow-up work using the facilitation strategies in the next section can examine the relationship between conversation dynamics and average turn length.

6 Defining participant sharing

Following the presentation of high-level computational measures of turn-taking in the corpus, we turn our attention to annotating speaker turns in the corpus according to several codebooks which we define. Because Fora conversation guides were designed to elicit personal stories and experiences from participants, we develop a codebook and annotation task related to identifying two types of sharing, personal story and personal experience, as the main outcomes of interest from the dialogue.

Personal story: Building on Antoniak et al. (2023), we define a personal story as a series of events involving the speaker that occurred in the past or in the ongoing active present.

Personal experience: Following a similar approach in Falk and Lapesa (2023) to defining "experiential knowledge" in contrast to story, we define personal experience as a mention of personal background, identity, fragmentary recall of events, or mentions of habitual or ongoing actions, occurrences, or speaker feelings within a speaker turn.

We choose speaker turns as our unit of labeling for the annotation task, and use the above definitions to apply labels to all speaker turns from a sample of 70 (26%) conversations from the corpus. To sample a subset of conversations from the corpus for annotation, we first sampled one from each of the 16 collections, then randomly sampled other conversations until we hit 25% of the corpus, or 66 conversations. We annotated 4 co-facilitated conversations as well, bringing the total number of annotated conversations to 70 conversations. We exclude co-facilitated conversations from model training and analysis on the facilitation strategy tasks due to potential differences in co-facilitated conversations that are not the focus of this work.

Speaker turns could be annotated with both labels if both kinds of personal sharing occurred within a turn. For example, this turn contains both: *My name's Addison. I'm from Machias in Washington County. And a time I felt like I mattered to my community was when my family was going through a hard time and some people from around our community sent presents on Christmas, and*

also helped with some of the bills. Addison's name and place of origin count as mentions of *personal experience*, while her story about going through a hard time, being sent presents, and being helped with bills qualifies as a *story* due to its described sequence of events.

6.1 Defining facilitation strategies

Facilitators employ a number of facilitation strategies in conversation, some of which are composite dialogue speech acts. Building on work from Chung (2011), we choose and define a subset of these for annotation, classification, and analysis. Examples and full definitions of each facilitation strategy are in the appendix.

1. Validation Strategies
   (a) **Express appreciation**
   (b) **Express agreement or affirmation**
2. Invitations to Participate
   (a) **Open invitation** to participants to participate
   (b) **Specific invitation to participate**, addressed to a particular speaker
3. Facilitation strategies
   (a) **Model examples** (facilitator models an example answer to the question or prompt)
   (b) **Follow-up question** (facilitator addresses a follow-up question or prompt to a speaker already speaking)
   (c) **Make connections** (facilitator makes a statement making a connection to or between different participants in the conversation, across topics in the conversation, or across participants and topics in the conversation collection.)

6.1.1 Annotation

We annotate participant sharing behaviors from 70 conversations across 10,625 total speaker turns. We solicited first-pass annotations from crowdworkers on the Prolific platform. We hosted an annotation server using Potato (Pei and Jurgens, 2023). After a short training and description of the task, workers were presented with 20 annotations. In the interface, workers could select one, both, or none of the participant sharing types for each speaker turn. Annotators had high recall in identifying personal sharing on the task, but sometimes misunderstood the nuanced distinction between story and experience. As such, if any two of the three annotators identified any type of sharing in the turn, one of three expert annotators reviewed the example to ascribe final labels to the sample.[2] These three expert annotators had very high (Krippendorff's alpha 0.81) agreement in a curated IRR task testing their understanding of all labels. The combination of these steps produced a high-quality expert-labeled subset of the data. We repeated this labeling procedure to annotate facilitation strategies, but subdivided the annotation task into three separate annotation questions, each of which was a multi-label question. (See appendix for additional details on crowdworker recruitment, task setup, and IRR). We used the resulting annotated subset as a gold-labeled subset of the data against which to evaluate some approaches to automate identification of these phenomena in the rest of the corpus.

6.2 Classifying participant sharing

Following work evaluating the suitability of language models for annotating corpora, (Pangakis et al., 2023; Savelka et al., 2023), we configure a prompt-based annotation pipeline using OpenAI's GPT-4 (Achiam et al., 2023) to generate annotations for this task under several conditions: zero shot, or few-shot with a list of curated example classifications, and with/without 2 previous turns of conversation context preceding the target turn. Following recent work showing the relatively better performance of fine-tuned models on nuanced tagging tasks compared to prompt-based methods (Thalken et al., 2023; Antoniak et al., 2023), we contrast prompt-based approaches to classifications from a single fine-tuned model. We use a 80/10/10 train/validation/test split for both tasks. We report our performance on the test set in Tables 1, 2 and 3.

7 Results

For each GPT-4[3] response, we extract the string representing the expected label in order to derive a binary value for the label for that sample. To evaluate performance, we treat the problem as a single-class classification problem for each task for maximum comparability across conditions. We fine-

[2] For each speaker turn, we also provided the two preceding turns as context to each annotator, given the potential importance of context to labeling decisions.

[3] We use model version gpt-4-1106-preview for all tasks.

Model	F1	Precision	Recall
	Story (n = 362)		
GPT-4, zero shot, w/ context	0.81	0.99	0.73
GPT-4, few shot, w/ context	0.81	0.87	0.77
GPT-4, zero shot	0.83	0.92	0.77
GPT-4, few shot	**0.86**	0.93	0.82
Fine-tuned RoBERTa	0.74	0.72	0.76
	Experience (n = 1367)		
GPT-4, zero shot, w/ context	0.71	0.73	0.70
GPT-4, few shot, w/ context	0.75	0.82	0.71
GPT-4, zero shot	0.75	0.75	0.75
GPT-4, few shot	0.75	0.78	0.72
Fine-tuned RoBERTa	**0.78**	0.75	0.80

Table 1: Personal sharing identification task

tune RoBERTa (Liu et al., 2019) (training details in the appendix) to perform a binary classification for each task. Table 1 shows model performance (using a macro-averaged f1 score for this and subsequent tasks) on the story and experience identification tasks. GPT-4 performed better on the *Personal story* identification task, perhaps because of the small class size (362 stories across the annotated 25% sample of the corpus), and we note high precision identifying this rare phenomenon with GPT-4. However, we note that fine-tuned RoBERTa's performance was better than GPT-4's on the task of identifying *Personal experience*, echoing Antoniak et al. (2023)'s finding that fine-tuning for nuanced tasks like these may perform better than prompt-based methods. However, our relatively lower performance than Antoniak et al. (2023) may be in part due to differences in transcribed speech and text written for online consumption.

7.1 Classifying facilitation strategies

We repeat a similar task setup for the list of seven facilitation strategies. Prompt-based methods were decomposed into three generation tasks: one each for "Validation" strategies, "Invitations to Participate," and "Facilitation strategies." Again, we fine-tune RoBERTa for a binarized version of each task and compare performance to results using the same four prompting styles with GPT-4.

Table 2 shows the model's performance on a single run of the *Appreciation*, *Affirmation*, *Open* and *Specific invitation* identification tasks. For all facilitation strategy tasks except *Follow up question*, GPT-4 performed better than fine-tuned RoBERTa. We note that GPT-4 errors on *Follow up question* often mistake *Open invitation* and *Specific invitation* for a *Follow up question*. The way we defined

Model	F1	Precision	Recall
	Appreciation (n = 827)		
GPT-4, zero shot, w/ context	**0.95**	0.93	0.97
GPT-4, few shot, w/ context	**0.95**	0.94	0.97
GPT-4, zero shot	**0.95**	0.94	0.97
GPT-4, few shot	**0.95**	0.94	0.95
Fine-tuned RoBERTa	0.93	0.88	1.0
	Affirmation (n = 320)		
GPT-4, zero shot, w/ context	**0.82**	0.84	0.81
GPT-4, few shot, w/ context	0.78	0.84	0.75
GPT-4, zero shot	0.81	0.85	0.77
GPT-4, few shot	0.77	0.84	0.72
Fine-tuned RoBERTa	0.72	0.75	0.69
	Open invite (n = 471)		
GPT-4, zero shot, w/ context	0.80	0.77	0.83
GPT-4, few shot, w/ context	**0.83**	0.82	0.85
GPT-4, zero shot	0.82	0.80	0.87
GPT-4, few shot	0.82	0.80	0.86
Fine-tuned RoBERTa	0.67	0.76	0.60
	Specific invite (n = 769)		
GPT-4, zero shot, w/ context	0.86	0.83	0.92
GPT-4, few shot, w/ context	**0.88**	0.85	0.92
GPT-4, zero shot	**0.88**	0.85	0.92
GPT-4, few shot	0.87	0.84	0.91
Fine-tuned RoBERTa	0.83	0.76	0.90

Table 2: Validation and invitation classification tasks

it, *Follow ups* were not necessarily "questions," so the word "question" in the class title might have skewed the GPT-4 results compared to our fine-tuning approach that learned more directly from the data. We also note that RoBERTa struggled to learn *Make connections* and *Provide example*, given the small class size and even smaller presence in the 10% validation/test sets. By comparison, GPT-4 performed decently well, suggesting that LLMs like GPT-4 may be useful in detecting rare phenomena with proper validation. We did not find that giving GPT-4 two turns of conversation context consistently improved performance.

7.2 Exploratory analysis

With what prevalence do we find each facilitation strategy in the corpus? We analyze the 10,625 facilitator speaker turns of gold-labeled data to understand prevalence, given that varied model baselines on constituent tasks could skew results if interpreted as predictions over the full corpus. *Appreciation* (n = 827) is the most common facilitation strategy we capture, and all conversations feature it at least once. *Affirmation* (n = 320) is also common, occurring in 91% of conversations. 93% of conversations feature at least one *Open invitation* (n = 471). *Specific invitations* (n = 769) were also common, featuring in 88% of all conversations. Closely related, *Follow ups* (n = 316) occur in 79%

Model	F1	Precision	Recall
Follow-up question (n = 316)			
GPT-4, zero shot, w/ context	0.63	0.61	0.79
GPT-4, few shot, w/ context	0.60	0.60	0.78
GPT-4, zero shot	0.61	0.59	0.70
GPT-4, few shot	0.60	0.58	0.68
Fine-tuned RoBERTa	**0.65**	0.61	0.70
Make connections (n = 102)			
GPT-4, Zero shot, w/ context	0.69	0.70	0.68
GPT-4, few shot, w/ context	0.69	0.76	0.65
GPT-4, zero shot	0.77	0.80	0.74
GPT-4, few shot	**0.78**	0.79	0.77
Fine-tuned RoBERTa	0.18	0.80	0.11
Provide example (n = 62)			
GPT-4, Zero shot, w/ context	0.68	0.64	0.74
GPT-4, few shot, w/ context	0.69	0.63	0.86
GPT-4, zero shot	0.65	0.61	0.74
GPT-4, few shot	**0.71**	0.65	0.87
Fine-tuned RoBERTa	0.44	0.85	0.30

Table 3: Full results of additional 3 facilitation strategies task, GPT-4 and RoBERTa

of conversations. We find *Make connections* (n = 102) in 71% of conversations, and though *Provide example* (n = 62) is quite rare in absolute terms, it still features at least once in 60% of conversations.

7.2.1 Facilitator sharing

In total, we find 362 personal stories and 1,367 mentions of personal experience in the 25% sample of the corpus. One stylistic choice facilitators make is the degree to which they share their own personal experiences in the conversation. 82% of conversations with a single facilitator feature at least one facilitator personal experience, and 90% of conversation participants share from personal experience. 31% of conversations feature at least one facilitator personal story, similar to the 34% of participants who share personal stories.

7.3 Exploratory analysis on facilitator behavior and participant sharing

Next, we explore if facilitation strategy relates to participant outcomes: story and experience sharing. Analyzing facilitated dialogue is complex, as dialogues may be analyzed for outcomes at the conversation, collection, participant, and facilitator levels. For this analysis, we define the rate of participant sharing as the percentage of participant speaker turns that contain participant sharing, and we use this as an outcome measure. We next test whether increased rates of facilitator *Open invitations* to share, and *Specific invitations* to a particular individual from the facilitator to share, correlate with the rates at which participants share. Using a Pearson correlation with Bonferroni correction to adjust for multiple comparisons, we find a statistically significant positive correlation between rates of participants sharing personal experiences within a conversation and rates of both open and specific facilitator invitations to share in a conversation ($p < .05$), but not between rates of participants sharing personal stories and either open or specific facilitator invitations to share.

We do not find a statistically significant correlation between conversation-level rates of participant and facilitator personal experience or story sharing. However, we do find a positive correlation ($p < .05$) between rates of facilitator *Affirmation* and participant *Personal story* sharing (but not *Personal experience* sharing), potentially speaking to positive facilitator response to stories shared. Similarly, we find facilitator rates of *Make connections* to be positively correlated ($p < .05$) with *Personal story* sharing, but not with *Personal experience* sharing. We find that *Make connections* and *Express affirmation* rarely directly precede or follow an instance of *Personal story*, so follow-up work is needed to better understand whether facilitators drive sharing through *Expressing affirmation* and *Making connections* or do them in response to sharing, as well as to understand what other variables may mediate participant sharing.

How does inequality in time-sharing relate to rates of participant sharing? To investigate, we return to the Gini coefficient as our measure of inequality in participant time and turn count sharing across a conversation. We fit linear regression models predicting a participant sharing (*Personal story* or *Personal experience*) rate in a conversation as a function of inequality (the Gini coefficient of either time or turn count from the conversation). We find a statistically significant negative relationship between inequality (both in terms of time and turn count) and rates of both personal story sharing and personal experience. This requires more investigation, and reasons for this relationship at a conversation level may differ. However, this finding may suggest facilitators should promote equitable time-sharing not just because it is kind to participants, but also because it is associated with a better conversation outcome.

Exploratory analyses like these, combined with high-confidence classifications that scale to the rest of the corpus, and methods that move beyond correlation, can help systematically make sense of Fora and other kinds of facilitated dialogue.

8 Research directions for the study of facilitated dialogue

We hope these data, measures, tasks, and analyses can encourage future work in this area, especially combined with behavioral experiments and other methods and insights across fields. We outline some areas of potential work:

- *Studying facilitators and participants:* How does facilitator behavior relate to participant outcomes and turn-taking behaviors as characterized above? How do facilitators change or maintain strategy across different conversations? How do participants and facilitators engage and respond to others in facilitated dialogue, mediated by personality, identity, and conversation format?

- *Conversation format:* How do the number of participants and facilitators and conversation timing goals affect facilitator behavior, conversation dynamics, and outcomes? How do online dialogues differ from in-person ones? How does a conversation guide or set of norms affect dynamics, and does the facilitator often adhere to the guide? Do different conversation guides affect outcomes, insights, and the range of discussed topics?

- *Understanding spoken facilitation:* Most datasets for NLP come from text-native data sources. How do transcribed stories and experiences differ from text-native stories and facilitated experiences, and how might we better develop tools to process them?

- *Dialogue as a method and intervention:* As facilitated dialogue increases in popularity as a method of civic engagement and social insight, how do we understand the dual role of dialogue as a potential method for insight as well as an intervention on participants' perception of self, group, and community? How might we measure dialogue quality through behavioral experiments, interviews, or other measures?

- *Dialogue agents:* LLMs mediate interviews with humans, but they are also increasingly positioned to mediate interactions across multiple humans. Are there tasks where LLMs may be suited to this style of interaction, and can this corpus be useful in responsibly assessing and/or developing them for appropriate use cases?

9 Conclusion & Future Work

We present a unique corpus of annotated facilitated conversations and a suggested framework for analyzing facilitated dialogue. We outline key tasks for understanding facilitated dialogue in terms of participant sharing behavior and facilitator behavior, create a 25% sample of human-labeled data, then establish baselines for each task. We find decent model performance on some tasks, and especially for rare phenomena, GPT-4 does decently well on these tasks in contrast to a fine-tuned RoBERTa model. In a data-poor context where fine-tuning is challenging, this points to the potential for large models to assist in making sense of small data, pending validation. Future work can investigate performance using open-source models on these tasks. Overall, we find that Fora conversations regularly elicit personal stories and experience. We find that some facilitation strategies are more common than others, and some strategies may relate to increased rates of participant personal sharing.

We believe this corpus, framework, and findings will fuel research on facilitated conversations among humans, effective facilitation of group dialogue, modeling goal-driven dialogue, understanding spoken stories in dialogue, and facilitation of groups of humans by language agents. We hope this work will serve as a starting point for a rich emerging science of facilitated dialogue.

Future work can compare conversation norms in facilitated dialogue to more traditional forms of interviewing, information collection, and conversation. Comparison datasets in the future could include podcast data or TV interviews with multiple conversationalists. In this work, we did not test every combination of prompting methods and training parameters, instead focusing on a high-level framework, definitions, and establishing baselines. There remains significant opportunity for improvement, including exploration of methods to improve finetuning for small classes. Future work can examine additional prompting strategies like randomly sampling from the training data to provide few-shot examples, as well as Chain-of-Thought prompting.

10 Limitations

This work lays out some key facilitation strategies we observe in the corpus, but this list of facilitation

strategies is nowhere near exhaustive. Understanding other facilitation strategies, like paraphrasing participant answers, requires follow-up work. Another limitation is that while this work focuses on spoken dialogue, the data is textual in nature, which introduces limitations regarding our ability to interpret non-linguistic context that is absent from the transcript. Some facilitation strategies, like holding silent space, also require analysis beyond the transcript, and the corpus timestamps could be useful for this purpose. Furthermore, the *Personal experience* category of sharing we defined in contrast to *Personal story* catches a range of different personal sharing behaviors, from a simple introduction to a discussion of habitual occurrences in a person's past. The range of behaviors captured within the category may be noisy for a model to capture, and may require refinement and deeper analysis to illuminate their significance to a conversation.

Findings are also represented here at the speaker turn level. Given the variation in turn lengths, annotations may need to be refined to a span for further research use depending on the desired use case. Furthermore, for stories that are collaboratively interrupted by others, our classification may give false negatives for some phenomena at the speaker turn level. This may systematically bias our analyses towards certain groups, depending on their norms regarding sharing time and interrupting. In this problem formulation, we thus may miss context that occurs across multiple speaker turns, or stories that are told across long spans of time with interruptions in between. We encourage future work examining this corpus using different units of analysis or identifying sharing and facilitation strategies at the span level and comparing findings to this work.

11 Ethics Statement

Participants in these conversations consented to being recorded and for their speaking turns to be included in a public dataset. Oftentimes, participants shared quite personal stories in these facilitated conversations. They were given the opportunity to redact any parts of the conversation they did not want shared. We are still taking precautions by de-identifying sensitive information from the corpus that will be available for research use.

Many participants in the Fora conversations were recruited because they come from disenfranchised communities whose voices are underrepresented in civic processes. This creates an opportunity for researchers to responsibly study the speech and stories that are shared in this unique corpus, but it also comes with responsibility. Stories from communities or disenfranchised groups could be misrepresented, causing unintended harm. We intend to make the corpus described in this project accessible through a form application, facilitating its use in research for positive impact but also protecting against misuse.

Acknowledgements

We thank Alex Kelly Berman, Kelly Paola, Vy Dao, and everyone involved in Cortico for their collaboration on this project. We thank the partner and community organizations involved in Cortico, LVN, and Real Talk for Change. We are grateful to Dimitra Dimitrakopoulou, as well as all other researchers and staff at MIT CCC for feedback and for making this work possible. We thank undergraduate research assistants including Grace Song, Irene Huang, Alison Wang, and Isha Agarwal for their valuable assistance at different stages of the project. We sincerely thank our anonymous reviewers for their thoughtful, engaged feedback as we improved this paper. We also acknowledge co-author Deb Roy's dual role as a professor at MIT and as unpaid CEO of Cortico at the time of publication of this paper.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Maria Antoniak, Joel Mire, Maarten Sap, Elliott Ash, and Andrew Piper. 2023. Where Do People Tell Stories Online? Story Detection Across Online Communities. ArXiv:2311.09675 [cs].

Chris Bail. 2022. *Breaking the social media prism: How to make our platforms less polarizing*. Princeton University Press.

Rebecca Bates, Patrick Menning, Elizabeth Willingham, and Chad Kuyper. 2005. Meeting acts: a labeling system for group interaction in meetings. In *Interspeech 2005*, pages 1589–1592. ISCA.

Laura W. Black. 2013. Framing Democracy and Conflict Through Storytelling in Deliberative Groups. *Journal of Deliberative Democracy*, 9(1).

David Bohm and Lee Nichol. 2004. *On dialogue*, routlege classics ed. edition. Routledge classics. Routledge, London ; New York. OCLC: ocm56368426.

Harry Bunt, Volha Petukhova, Andrei Malchanau, Alex Fang, and Kars Wijnhoven. 2019. The DialogBank: dialogues with interoperable annotations. *Language Resources and Evaluation*, 53(2):213–249.

Wing Sum Cheung and Khe Foon Hew. 2010. Examining facilitators' habits of mind in an asynchronous online discussion environment: A two cases study. *Australasian Journal of Educational Technology*, 26(1).

LIM Sze Chung. 2011. Critical Thinking in Asynchronous Online Discussion: An Investigation of Student Facilitation Techniques.

Mark G Core and James F Allen. 1997. Coding Dialogs with the DAMSL Annotation Scheme.

Kara N. Dillard. 2013. Envisioning the Role of Facilitation in Public Deliberation. *Journal of Applied Communication Research*, 41(3):217–235. Publisher: Routledge _eprint: https://doi.org/10.1080/00909882.2013.826813.

Katharina Esau. 2018. Capturing Citizens' Values: On the Role of Narratives and Emotions in Digital Participation. *Analyse und Kritik*, 40.

Neele Falk and Gabriella Lapesa. 2023. StoryARG: a corpus of narratives and personal experiences in argumentative texts. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2350–2372, Toronto, Canada. Association for Computational Linguistics.

James S. Fishkin. 1997. *The Voice of the People: Public Opinion and Democracy*. Yale University Press.

Marlène Gerber, André Bächtiger, Susumu Shikano, Simon Reber, and Samuel Rohr. 2018. Deliberative Abilities and Influence in a Transnational Deliberative Poll (EuroPolis). *British Journal of Political Science*, 48(4):1093–1118. Publisher: Cambridge University Press.

Shazia Akbar Ghilzai. 2015. Conversational Analysis of Turn taking Behavior and Gender Differences in Multimodal Conversation.

J. J. Godfrey, E. C. Holliman, and J. McDaniel. 1992. SWITCHBOARD: telephone speech corpus for research and development. pages 517–520. IEEE Computer Society.

Soomin Kim, Jinsu Eun, Joseph Seering, and Joonhwan Lee. 2021. Moderator Chatbot for Deliberative Discussion: Effects of Discussion Structure and Discussant Facilitation. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1):1–26.

Judith A. Kolb, Sungmi Jin, and Ji Hoon Song. 2008. A model of small group facilitator competencies. *Performance Improvement Quarterly*, 21(2):119–133. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/piq.20026.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. ArXiv:1907.11692 [cs].

Derek Matravers and Jonathan Pike. 2005. *Debates in Contemporary Political Philosophy: An Anthology*. Routledge. Google-Books-ID: p_FlHTD3ZmgC.

Mizuki Motozawa, Yohei Murakami, Mondheera Pituxcoosuvarn, Toshiyuki Takasaki, and Yumiko Mori. 2021. Conversation Analysis for Facilitation in Children's Intercultural Collaboration. In *Proceedings of the 20th Annual ACM Interaction Design and Children Conference*, IDC '21, pages 62–68, New York, NY, USA. Association for Computing Machinery.

Anthony John Onwuegbuzie, Wendy B. Dickinson, Nancy L. Leech, and Annmarie G. Zoran. 2009. Toward More Rigor in Focus Group Research: A New Framework for Collecting and Analyzing Focus Group Data. *International Journal of Qualitative Methods: ARCHIVE*, 8(3):1–21. Number: 3.

Nicholas Pangakis, Samuel Wolken, and Neil Fasching. 2023. Automated Annotation with Generative AI Requires Validation. ArXiv:2306.00176 [cs].

Jiaxin Pei and David Jurgens. 2023. When do annotator demographics matter? measuring the influence of annotator demographics with the POPQUORN dataset. In *Proceedings of the 17th Linguistic Annotation Workshop (LAW-XVII)*, pages 252–265, Toronto, Canada. Association for Computational Linguistics.

Andrew Reece, Gus Cooney, Peter Bull, Christine Chung, Bryn Dawson, Casey Fitzpatrick, Tamara Glazer, Dean Knox, Alex Liebscher, and Sebastian Marin. 2023. The candor corpus: Insights from a large multimodal dataset of naturalistic conversation. *Science Advances*, 9(13):eadf3197.

David M. Ryfe. 2006. Narrative and Deliberation in Small Group Forums. *Journal of Applied Communication Research*, 34(1):72–93.

H. Saunders. 1999. *A Public Peace Process: Sustained Dialogue to Transform Racial and Ethnic Conflicts*. Springer. Google-Books-ID: zKJ9DAAAQBAJ.

Jaromir Savelka, Kevin D. Ashley, Morgan A. Gray, Hannes Westermann, and Huihui Xu. 2023. Can GPT-4 Support Analysis of Textual Data in Tasks Requiring Highly Specialized Domain Expertise? In *Proceedings of the 2023 Conference on Innovation and Technology in Computer Science Education V. 1*, pages 117–123. ArXiv:2306.13906 [cs].

Charlotte Schluger, Jonathan P. Chang, Cristian Danescu-Niculescu-Mizil, and Karen Levy. 2022. Proactive moderation of online discussions: Existing practices and the potential for algorithmic support. *Proc. ACM Hum.-Comput. Interact.*, 6(CSCW2).

Tsukasa Shiota, Takashi Yamamura, and Kazutaka Shimada. 2018. Analysis of Facilitators' Behaviors in Multi-party Conversations for Constructing a Digital Facilitator System. In *Collaboration Technologies and Social Computing*, Lecture Notes in Computer Science, pages 145–158, Cham. Springer International Publishing.

David W. Stewart and Prem N. Shamdasani. 2014. *Focus Groups: Theory and Practice*. SAGE Publications. Google-Books-ID: 1svuAwAAQBAJ.

Tanya Stivers, N. J. Enfield, Penelope Brown, Christina Englert, Makoto Hayashi, Trine Heinemann, Gertie Hoymann, Federico Rossano, Jan Peter de Ruiter, Kyung-Eun Yoon, and Stephen C. Levinson. 2009. Universals and cultural variation in turn-taking in conversation. *Proceedings of the National Academy of Sciences*, 106(26):10587–10592. Publisher: Proceedings of the National Academy of Sciences.

Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. 2000. Dialogue Act Modeling for Automatic Tagging and Recognition of Conversational Speech. *Computational Linguistics*, 26(3):339–373.

Rosamond Thalken, Edward Stiglitz, David Mimno, and Matthew Wilkens. 2023. Modeling legal reasoning: LM annotation at the edge of human agreement. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9252–9265, Singapore. Association for Computational Linguistics.

Thomas P. Wilson, John M. Wiemann, and Don H. Zimmerman. 1984. Models of Turn Taking in Conversational Interaction. *Journal of Language and Social Psychology*, 3(3):159–183. Publisher: SAGE Publications Inc.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Ziang Xiao, Michelle X. Zhou, Wenxi Chen, Huahai Yang, and Changyan Chi. 2020. If i hear you correctly: Building and evaluating interview chatbots with active listening skills. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–14, New York, NY, USA. Association for Computing Machinery.

A Appendix

A.1 Conversation lengths

In Figure 3, one can see the spread of durations of Fora conversations in the data set.

A.2 Personal story and experience

Longer definitions and examples:

A **personal story** describes a discrete sequence of events involving the speaker that occurred in the past or in the ongoing active present.

The mention of **personal experience** includes mentions of name, facts about self, professional background, or general statements about the speaker's life that are not a sequence of specific events.

Example 1

CONTEXT TURN: I'm ready to go next.
CONTEXT TURN: Ok, go ahead.
TARGET TURN: I started in vegetables. And then when I came to Maine, and I looked around, and I said, "Jesus, every year there's like 12 more vegetable farms, but I can only go one place to get a steak or some sausage. Wait a minute here. There's a niche." And that's what I think they have to be open to. So I switched completely.

Answer: Personal story

Example 2

CONTEXT TURN: Can I tell you something?
CONTEXT TURN: Go on.
TARGET TURN: In my 15 years of teaching, I rarely saw students behave badly to each other.

Answer: Personal experience

Example 3

CONTEXT TURN: I'm ready to go next.
CONTEXT TURN: Ok, go ahead.
TARGET TURN: My favorite thing is that I grew up in a town with 16 000 people so this way bigger than anything I've ever lived in. Unfortunately for me I've only lived here for like a year and a half, almost two years. So I moved right fresh out of college and into a pandemic basically, and so I guess I haven't seen all of the great things in Madison does, but this is the biggest, and I felt very welcomed in this area so that's that.

Answer: Personal story, Personal experience

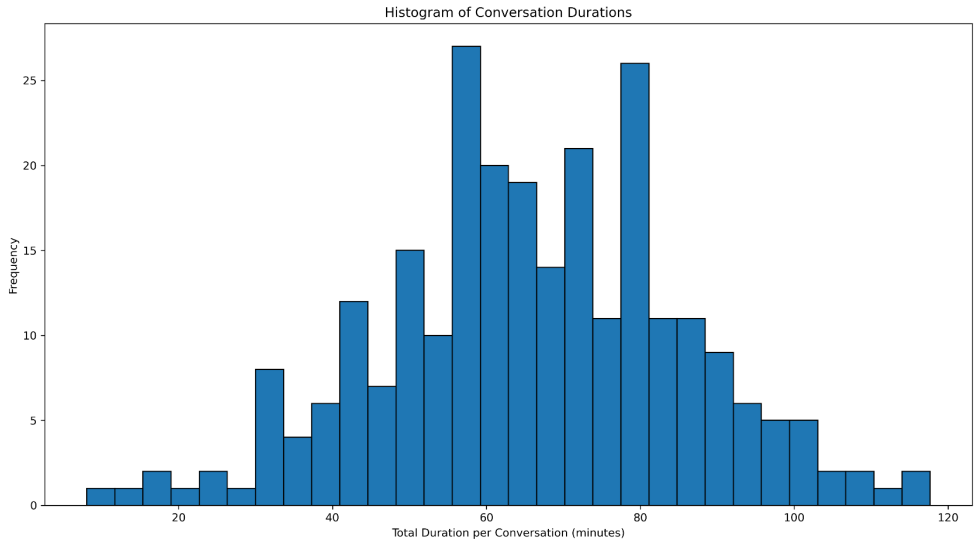


Figure 3: Duration of Fora Conversations

Example 4
CONTEXT TURN: Wendy, you're on mute now.
CONTEXT TURN: Ok, sorry, but it's your turn.
TARGET TURN: No, I thought it was your turn? Sorry, I'm confused.
Answer: Neither

A.3 List of facilitator dialogue acts and definitions

1. Validation Strategies
   (a) **Express appreciation**
      i. "Thanks for sharing that."
      ii. "Alright, great, now do I have everyone? Appreciated. Sure. Let's move on."
   (b) **Express agreement or affirmation**
      i. "Definitely. I hear you."
      ii. "I agree. I keep hearing that in these conversations."
2. Invitations to Participate
   (a) **Open invitation** to participants to participate
      i. "Alright with that being said, does anyone want to kick us off?"
      ii. "Ok great. Now anyone who wants to go next can unmute."
   (b) **Specific invitation to participate**, addressed to a particular speaker
      i. "Kelly? Did you want to go next?"
      ii. "Great. That's all I wanted to put on the table. Oh, Joe? I see you're raising your hand. Are you ready to go?"
3. Facilitation strategies
   (a) **Model examples** (facilitator models an example answer to the question or prompt)
      i. "So first we are going to give introductions and then share a value that's important to us. I'll kick us off. I'm Kelly and a value that's important to me is courage."
      ii. "Sharing stories is what we're doing next. Some people are confused in this stage so I'll give you an example. My grandmother lived next door to us when I was growing up. One time, the sheriff came by, and I was playing outside. It scared me to see this guy, an unknown man, come to our house when we lived that far out in the country. And my grandmother's attitude towards him didn't help. Since then I've been scared of the police. Ok does anyone want to share your first encounter with police next?"
   (b) **Follow-up question** (facilitator addresses a follow-up question to a speaker

Collection Title	Size
Maine ED 2050	76
United Way of Dane County	43
NYC Community Voices: COVID-19 & Beyond	22
Public Narrative	19
EGI Conversations	19
Cambridge City Manager Selection Project	18
Virtual Open Newsroom	13
Engage 2020	8
Cap Times People's Agenda	6
Digital Coverage of Crime & Justice Conversations	6
Foundation for Black Women's Wellness	6
LPA Yo Soy Conversations	6
Census 2020 Conversations	5
Equitable Economies Dialogues	5
Police & Fire Commission Conversations	5
Santa Cruz Local	5

Table 4: Size of each collection within Fora corpus

already speaking)

i. "Wow, Kelly, that's such a touching story. Did you ever find out who sent the package?"

(c) **Make connections** (facilitator includes a statement making a connection to or between different participants in the conversation, or across topics in the conversation)

i. "That's similar to what Jamie shared earlier, right?"

ii. "I have a similar experience that resonates."

iii. "That connects to what we were discussing earlier."

A.4 Corpus meta-data

Conversation collections in the corpus were convened for diverse reasons in a variety of locations.

Example conversation collections include:

- **United Way of Dane County**. A collection of 43 public conversations hosted by 25 community stakeholders in Dane County, Wisconsin to celebrate the 100th anniversary of the community organization and inform planning for the next 100 years.
- **Cambridge City Manager Selection Project**. A collection of 18 public conversations hosted in advance of a City Manager Selection Project in Cambridge, MA, ultimately conducted by 9 members of the Cambridge City Council. Feedback from the conversations was used to inform a leadership profile outlining the role's criteria in advance of selection.
- **NYC Community Voices: COVID-19 and Beyond**. A collection of 22 public conversations hosted in partnership with the NYC Department of Health & Mental Hygiene, the NYC Public Health Corps (PHC), 13 PHC community-based organization partners to engage residents in sharing their stories. 100+ people from across 35 historically underserved NYC neighborhoods were recruited.

We provide conversation-level metadata including information about:

- Location (partial metadata)
- Date of conversation
- Conversation duration
- Audio offsets for each conversation turn

Collection-level metadata is also partially available for each conversation, including:

- Conversation guide (document)
- Sensemaking reports and websites, when applicable
- Description of collection context

All 16 conversation collections in this corpus verbally collected consent to conversation recording and public sharing, and for the use of the recordings in research. Recruitment for most collections was done through participating community partners and organizations. Most conversation collections recruited on a volunteer basis. Two collections, NYC Community Voices: COVID-19 & EGI Conversations, compensated participants in the conversations.

A.5 Prolific annotations & IRR

Prolific workers were recruited for two distinct phases of annotation: one task for *Personal story* and *Personal experience* and the other for the seven identified *Facilitation strategies*. Prolific workers were given 20 annotations each, took an average of 20 minutes for each task, and were compensated at an average $13.50 USD per hour. Workers were recruited from the US and UK, with 95%+ task approval rating, first language of English, and an education level of at least some college. We sought and

received IRB exempt approval for the surveys annotators took. For the *Personal story* and *Personal experience* task, annotators saw a single question with checkboxes to denote the presence of either one. In the Facilitation strategies task, annotators saw three questions per conversation turn, each requesting check-box style multilabels for *Validation Strategies* (*Express affirmation*, *Express appreciation*) *Invitations to Participate* (*Open invitation* and *Specific invitation*), and the other three *Facilitation strategies* (*Make connections*, *Follow up question*, *Provide examples*).

On a curated, class-balanced subset of the data of 27 examples with at least 5 examples of each facilitation strategy, we ran an inter-annotator agreement activity with 19 annotators from our worker pool. Using Krippendorff's Alpha as a measure of inter-annotator agreement, the tasks range in their level of agreement. For "Express appreciation," "Open invitation," and "Specific invitation," we achieve moderate levels of agreement in the .4-.48 range. For "Express agreement," "Make connections," "Provide example," and "Follow-up question" we achieve poor to fair levels of agreement but high recall. However, expert annotators on our research team achieved an agreement level of .81, or very high levels of agreement on the task. As such, expert annotators reviewed annotator-suggested labels to construct a high-confidence final set of annotations.

A.6 Additional visualizations

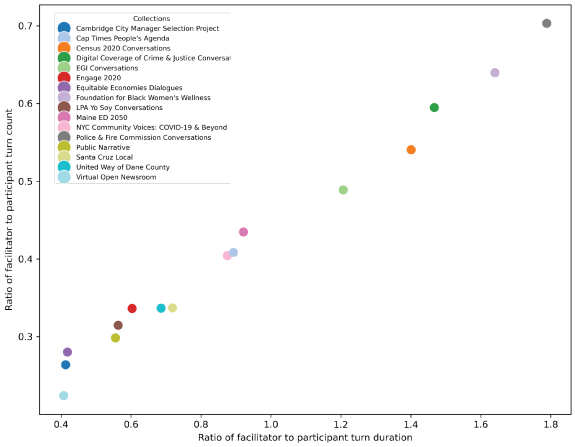


Figure 4: Turn-taking in Fora Conversations

Figure 4 plots the ratio of facilitator to participant turn duration versus ratio of facilitator to participant turn count. The collections represent a range of values in participant & facilitator turn-taking ratios, which may be used in future analysis to explain facilitator and participant behavior in the conversations.

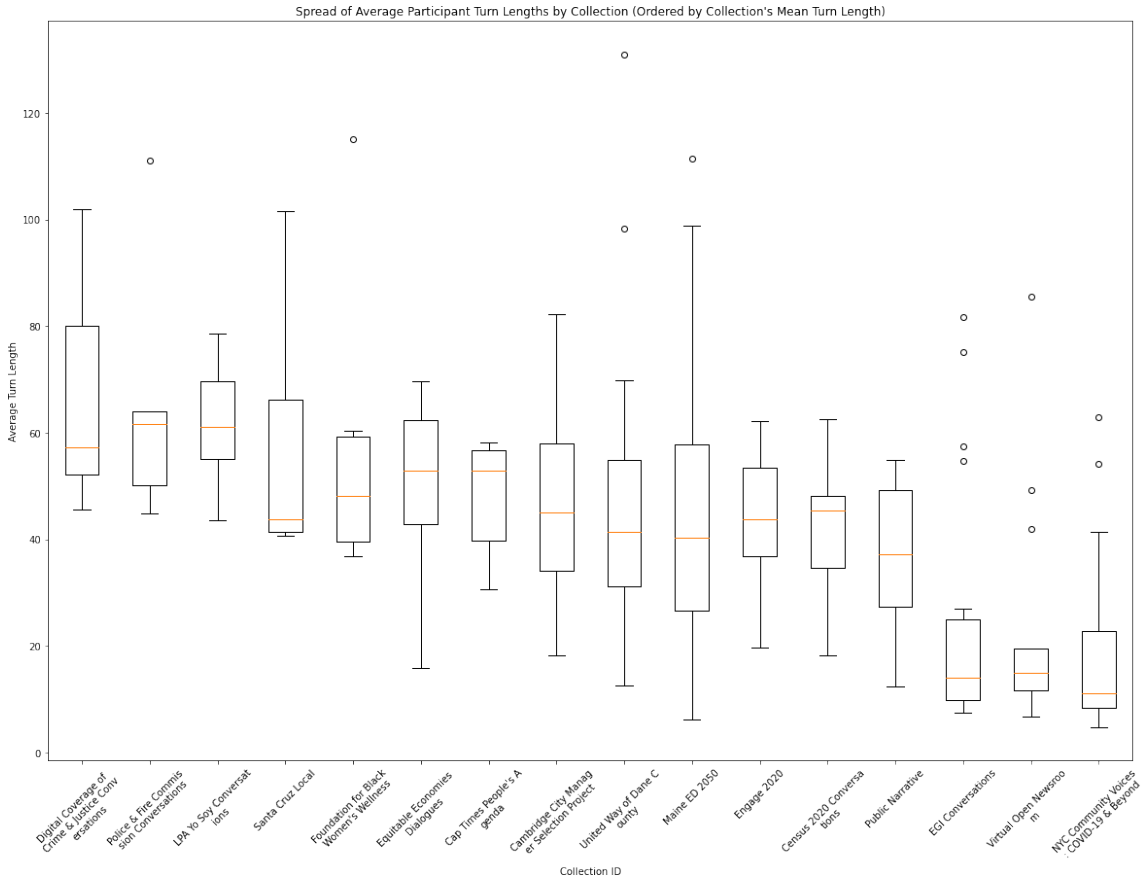


Figure 5: Average turn length across Fora collections differs significantly. Longer average turn length suggests a facilitation style more akin to a sequence of monologues, while shorter average turn length suggests a facilitation style more akin to casual conversation.

A.7 Additional model details

Fine-tuning RoBERTa: We use an 80/10/10 train/validation/test split on the human-annotated corpus of 66 conversations for facilitation strategies and 71 conversations for personal sharing. We train using a HuggingFace (Wolf et al., 2020) instantiation of a pre-trained RoBERTa model (`'roberta-base'`).

A.8 Prompts

We use the following prompt templates to prompt GPT-4 (temperature of 0) in 4 tasks: one for *Personal sharing* and *Personal story*, one for *Validation strategies*, one for *Invitations to participate*, and one for *Facilitation strategies*.

A.8.1 Personal sharing - prompting example

We next explain our prompt construction for the task identifying *Personal sharing* and *Personal story*.

For conditions that did gave 2 context turns, the first instruction was:

"You will be presented with text in a TARGET TURN from a speaker turn from a transcribed spoken conversation. The text was spoken by a participant in a conversation. We are identifying instances of personal sharing by the speaker. Your job is to identify the following sharing types in the quote:"

For conditions that did not give 2 context turns, the first instruction was:

"You will be presented with text from a speaker turn from a transcribed spoken conversation. The text was spoken by a participant in a conversation. We are identifying instances of personal sharing by the speaker. Your job is to identify

the following sharing types in the quote:"

Then, the definitions were given:

"- Personal story

- Personal experience

These are types of sharing that are only sometimes used. Many of the quotes will not contain either, and some will contain both. The definitions are important to make sure they actually apply.

Definitions:

Personal story: A personal story describes a discrete sequence of events involving the speaker that occurred at a discrete point in time in the past or in the ongoing active present.

Personal experience: The mention of personal experience includes introductions, facts about self, professional background, general statements about the speaker's life that are not a sequence of specific events, and repeated habitual occurrences or general experiences that are not discrete."

Then, for few-shot conditions, examples were appended to the prompt. For conditions without context, the references to "CONTEXT" and "TARGET" turns were deleted.

"Here are some examples and expected correct answers for each:

Example 1

CONTEXT TURN: I'm ready to go next.

CONTEXT TURN: Ok, go ahead.

TARGET TURN: I started in vegetables. And then when I came to Maine, and I looked around, and I said, "Jesus, every year there's like 12 more vegetable farms, but I can only go one place to get a steak or some sausage. Wait a minute here. There's a niche." And that's what I think they have to be open to. So I switched completely.

Answer: Personal story

Example 2

CONTEXT TURN: You know what's crazy?

CONTEXT TURN: Go on.

TARGET TURN: In my 15 years of teaching, I rarely saw students behave badly to each other.

Answer: Personal experience

Example 3

CONTEXT TURN: I'm ready to go next.

CONTEXT TURN: Ok, go ahead.

TARGET TURN: My favorite thing is that I grew up in a town with 16,000 people so this way bigger than anything I've ever lived in. Unfortunately for me I've only lived here for like a year and a half, almost two years. So I moved right fresh out of college and into a pandemic basically, and so I guess I haven't seen all of the great things in Madison does, but this is the biggest, and I felt very welcomed in this area so that's that.

Answer: Personal story, Personal experience

Example 4

CONTEXT TURN: Wendy, you're on mute now.

CONTEXT TURN: Ok, sorry, but it's your turn.

TARGET TURN: No, I thought it was your turn? Sorry, I'm confused.

Answer: None"

Then, the final instruction was given. For the condition without conversation context, it was: "Annotate the speaker turn for the above personal sharing types in conversation. Do not output any explanation, just output a comma-separated list of any that apply. If none apply, output "None".

Answer:"

For the condition containing context, this last instruction was:

"Annotate the TARGET TURN for the above personal sharing types in conversation. Do not output any explanation, just output a comma-separated list of any that apply. If none apply, output "None".

Answer:"

All of these sub-parts of the prompt were concatenated together into a single prompt. This same prompt construction template was used to construct prompts for *Validation strategies*, *Invitations to participate*, and *Facilitation strategies*, with definitions and examples inserted accordingly.