# Computational Analysis of Conversation Dynamics through Participant Responsivity

**Margaret Hughes** **Brandon Roy** **Elinor Poole-Dayan** **Deb Roy** **Jad Kabbara**
MIT Center for Constructive Communication
Massachusetts Institute of Technology
{mhughes4, bcroy, elinorpd, jkabbara, dkroy}@mit.edu

## Abstract

Growing literature explores toxicity and polarization in discourse, with comparatively less work on characterizing what makes dialogue prosocial and constructive. We explore conversational discourse and investigate a method for characterizing its quality built upon the notion of "responsivity"—whether one person's conversational turn is responding to a preceding turn. We develop and evaluate methods for quantifying responsivity—first through semantic similarity of speaker turns, and second by leveraging state-of-the-art large language models (LLMs) to identify the relation between two speaker turns. We evaluate both methods against a ground truth set of human-annotated conversations. Furthermore, selecting the better performing LLM-based approach, we characterize the nature of the response—whether it responded to that preceding turn in a substantive way or not. We view these responsivity links as a fundamental aspect of dialogue but note that conversations can exhibit significantly different responsivity structures. Accordingly, we then develop conversation-level derived metrics to address various aspects of conversational discourse. We use these derived metrics to explore other conversations and show that they support meaningful characterizations and differentiations across a diverse collection of conversations.

## 1 Introduction

Trust in government is decreasing rapidly while political polarization increases. The toxicity that is pervasive in social media platforms like Twitter/X has seeped into our engagement offline. Town halls, community forums, and various other means of civic participation have grown hostile and unproductive (Innes and Booher, 2004; Tracy and Durfy, 2007). Democracy scholars call for systems that improve the health of the public sphere and for avenues that enable civic agency and dignity (Allen, 2023). Such systems enable citizens to gather to discuss meaningful ideas, work together to develop a shared understanding, and potentially even reach consensus in decision making processes. One example is citizens assemblies where groups selected through sortition gather together to learn, deliberate, and develop recommendations to their governing body based on the needs and goals of their community through small group, facilitated conversations (Chwalisz, 2019, 2020). These conversations can surface insights that prove valuable not only as a mirror to one's own community, but also as a portal into the thoughts and needs of a group for leadership or outsiders.

Growing literature explores toxicity, polarization, and decreased liberties within discourse, and while this understanding is important, that is only half of the challenge. Aspirationally, we strive not just for neutral discourse spaces, but actively constructive, healthy, and rich communication spaces. But how do we evaluate the quality of discourse with respect to these goals? We draw inspiration from collaboration literature and facilitated dialogue practice and argue that within a conversation, one fundamental ingredient for the constructiveness of conversation is *responsivity*: the extent to which participants in a dialogue actively listen to, respond to, and build upon one another. To understand this behavior in conversations, we operationalize and evaluate responsivity as a conversation quality metric. In simple terms, responsivity captures whether one person's conversational turn is responding to a preceding turn.

We develop and evaluate methods for quantifying responsivity—first through semantic similarity of speaker turns, and second by leveraging state-of-the-art large language models (LLMs) to compare the relation between two speaker turns. We evaluate both methods against a ground truth set of human-annotated conversations. Selecting the better performing LLM-based approach, we characterize the nature of the response—whether it was



*Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 35512–35531
November 4-9, 2025 ©2025 Association for Computational Linguistics

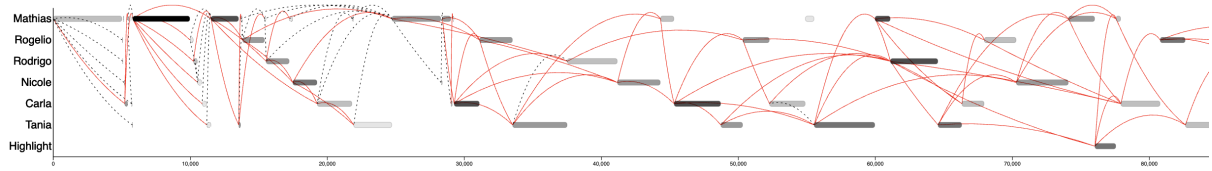


Figure 1: Conversation Map showing the flow of turns, their sequence, and the responsivity links.

responding to that preceding turn in a substantive way or not.

While we view responsivity links as a fundamental aspect of dialogue, we note that conversations can exhibit significantly different responsivity structure. Accordingly, we develop conversation-level derived metrics as a lens through which to examine different aspects of conversational discourse. We use these derived metrics to explore different kinds of conversations and show that they support meaningful characterizations and differentiations (that align with actual differences in purpose, style, etc.) across a diverse collection of conversations.

## 2 Related Work

### 2.1 Facilitated Dialogue

Facilitated dialogue is a conversation structure in which a facilitator guides participants in having a conversation according to a pre-designed conversation guide. Facilitators act as neutral stewards of the conversation to ensure conversation norms are respected, intervening in cases of norm violation. In the conversation space we focus on, one additional goal for facilitation is to encourage participants to share personal experiences rather than opinions. The goal of these conversations is thus not to persuade others or win through argumentation, but rather to understand one another more deeply through sharing of personal experiences. This approach is taken within conflict resolution and for community building and can be used to facilitate empathy, understanding, and connection among participants, within and across divides.

Further, facilitated dialogue is a technique used within deliberation and civic discourse spaces such as Citizens' Assemblies (Chwalisz, 2019, 2020) which consist of a random sample of people from a constituency gathered together to deliberate on specific social and political issues. In Citizens' Assemblies and other deliberative spaces, dialogue is the primary means of participation in the public sphere and empowers the various civic actors to practice agency and participate in their governance.

The concept of responsivity builds upon established theoretical frameworks in dialogue studies and communication theory. Drawing from Bakhtin (2010)'s notion of dialogic responsiveness, where meaning emerges through the dynamic interaction between speakers and listeners, our operationalization of responsivity captures the fundamental dialogic principle that authentic communication requires active engagement with others' contributions rather than mere sequential turn-taking.

This theoretical foundation aligns with research in conversational analysis and discourse studies that emphasizes the collaborative nature of meaning-making in dialogue (Schegloff, 2007). By quantifying the extent to which participants build upon, reflect back, and meaningfully engage with preceding contributions, our metrics operationalize key principles from facilitated dialogue practice where the quality of listening and responding directly impacts the depth and constructiveness of collective understanding.

The distinction between substantive and mechanical responsivity further reflects established practices in dialogue facilitation, where facilitators distinguish between responses that advance collective understanding and those that merely maintain conversational flow without adding substantive content (Isaacs, 1999). This theoretical grounding situates our computational approach within broader scholarly understanding of what constitutes effective dialogue and meaningful human interaction.

### 2.2 Conversation Metrics

While tools like Jigsaw's Perspective API (Lees et al., 2022) has been used widely to evaluate dynamics within online conversation spaces (Choi et al., 2015; Saveski et al., 2021), rarely are these metrics developed relationally–each comment or Tweet is generally treated independently of those that came before. Interactions and relationships between people are not the unit of analysis, but rather the text in isolation.

Recent work has started to explore what makes conversations pro-social and constructive. Bao

et al. (2021) look at turn-specific measures to identify pro-social conversation dynamics on Reddit. These metrics observe individual conversation contributions rather than the interwoven dynamics that emerge from participants responding to and being in relationship with one another.

Dowell et al. (2019) develop a system of metrics exploring the relationship between conversation participants. They present a core metric, namely responsivity, or the tendency of an individual to respond (or not) to the contributions of their collaborative peers. Responsivity is calculated based on the cosine distance between conversation turn embeddings – the closer the embeddings, the more responsive participants are to one another. We are inspired by this work for our own metric development and evaluation. We believe the choice to center relationship and context when evaluating and describing conversation aligns with the core values and practices in facilitated dialogue.

Another closely related line of work is conversation disentanglement, which seeks to separate interleaved conversational threads in multi-party dialogue (Zhu et al., 2021). Disentanglement methods aim to recover the underlying "who-responds-to-whom" structure, often in chaotic or asynchronous contexts such as IRC channels (Kummerfeld et al., 2019). More recent work has extended this to novel domains such as scripted or dramatic dialogue (Chang et al., 2023). Our approach builds on this tradition of mapping conversational links but diverges in its aims: while disentanglement focuses on thread recovery, we distinguish types of response (mechanical vs substantive) and develop conversation-level metrics to characterize the quality of dialogue. In this sense, our work complements disentanglement by enriching structural mappings with relational measures of responsiveness and constructiveness.

### 2.3 LLMs for Social Science

Recent advances in NLP have enabled more widespread use of LLMs in social science settings. LLMs have been applied to analyze social dynamics in several ways, including understanding emotional undertones (Dutt et al., 2024), social stances in conversations online (Chae and Davidson, 2023), as well as to extract speaker characteristics (Jurafsky et al., 2009; Broniatowski, 2012). In particular, leveraging LLMs as zero or few-shot annotators has been shown to be extremely promising (Gilardi et al., 2023; Wang et al., 2021; Ding et al., 2023; He et al., 2024; Huang et al., 2023), potentially even for subjective, nuanced tasks (Ziems et al., 2024; Ruckdeschel, 2025; Xiao et al., 2023). This may open up NLP research to tackle more complex, interdisciplinary, or niche datasets for which human annotation is very difficult or expensive (Ruckdeschel, 2025).

However, there is concern that LLMs trained on synthetic data may struggle on highly subjective tasks (Li et al., 2023). More work has shown that LLM annotation performance may struggle with conversational data (Ziems et al., 2024) and that models can be highly variable to prompts (Atreja et al., 2024). Some studies suggest that practitioners should use caution when using LLMs to annotate data (Pangakis et al., 2023; Huang et al., 2023). Pangakis et al. argue that the use of LLMs to automate annotation for research must always validate performance against human-annotated labels. Motivated by this, our methodology follows the best practices outlined by previous works.

Further, we build upon a growing literature that uses LLMs as a means to annotate and understand discussions at a large scale – a task previously quite inaccessible. We look to Korre et al. (2025) who do an overarching survey of how LLMs are used to evaluate and facilitate conversation, especially in digital domains like Reddit. We build upon their work by including key features such as turn taking in our analysis, but expand it further to pay particular attention to responsivity, or response relationships between participants, as a critical component of dialogue. Others explore methodologically various means to apply these metrics and evaluate their accuracy through mixed-methods approaches using LLMs to measure constructiveness in conversations, finding LLMs and hybrid-LLM approaches effective for the task (Zhou et al., 2024). A great deal of work further explores means of evaluation discourse and deliberation through descriptive metrics such as open-mindedness, equality of participation, a general respect for others, or progress towards a common goal (Barrett et al., 2024; Ercan et al., 2022). Yet, a gap continues to persist around responsiveness and connection to other participants.

## 3 Responsivity

As dialogue is about connection between participants, we explore how participants actively listen to, build upon, and reflect back contributions

of those before them through responsivity. While responsivity is not the only metric of importance when understanding conversation dynamics or quality, it is an important and understudied component of conversations, so we start with it in our work.

In studying responsivity, we define our unit of focus to be a conversation turn. This is the contiguous sequence of utterances a conversation participant speaks until another speaker starts their turn. Previous work considers responsivity between participants over a whole conversation (Dowell et al., 2019), while we calculate responsivity between participants across windows within a conversation to observe more granular, concurrent interactions. This granularity enables us to not only explore one conversation in summary, but to identify moments of conversations that yielded higher or lower responsivity, or highlighted relationships between participants at key interactions. We further explore the concept of responsivity by distinguishing between two kinds of responsivity:

- **Substantive responsivity:** An interaction where one person meaningfully engages with what another has said. It captures how much a speaker reflects back, builds upon, inquires about, or connects to other ideas, emotions, or experiences shared by the previous speaker, or answers a meaningful question from a previous speaker.
- **Mechanical responsivity:** An interaction that occurs when a speaker responds in a way that acknowledges or moves the conversation forward but does not add substantial new content. These responses may include polite phrases, conversational hand-offs, or social cues.

Responsive structures are integral to many forms of human interaction and conversation. As such, there is a need to better define and identify the boundaries of constructive communication. However, there is no existing method to map a conversation structure to understand how participants respond to and build upon one another. Furthermore, for humans to annotate a conversation for these turns is inefficient, especially if one might want to iterate upon a conversation design based on the responsivity within a previous conversation quickly, or if one would like to understand dynamics in a large corpus. Therefore, we ask how we might automate responsivity annotation. We describe an initial set of automation methods in the following section, along with the methods used to evaluate those approaches.

## 4 Methods

To automate annotating responsivity, we explore semantic similarity metrics and the use of LLMs via prompting. We develop a crowdsourced human annotation task to evaluate the automated methods and design an interactive data visualization.

### 4.1 Semantic Similarity

One approach to operationalizing responsivity is through *semantic similarity*, motivated by the idea that the content of the response should have some semantic overlap with the turn to which it is responding. To compute semantic similarity between conversation turns $i$ and $j$, we first obtain the sentence embedding of each conversation turn using MPNet (Song et al., 2020), a deep-learning based embedding model.[1] We then compute the cosine distance between the embedding vectors for turns $i$ and $j$. For a given turn, we compute its cosine similarity to the preceding 10 turns, and form a responsivity links to those turns with responsivity above a threshold.

### 4.2 LLM Approach

Large language models have shown remarkable performance across many tasks that involve certain kinds of reasoning, content analysis, and the generation and synthesis of text. We hypothesize that an LLM might be able to interpret the meanings in conversation turns in a more nuanced way compared to semantic similarity based on the MPNet embedding model.

We use two state-of-the-art LLMs: GPT-4o (OpenAI, 2024) and Claude 3.5 Sonnet (Anthropic, 2024)[2] to carry out three increasingly fine-grained tasks, the first of which is equivalent to the semantic similarity task introduced in Section 4.1.

**Stage 1 (turn-level linking):** Given a speaker turn and the 10 preceding conversation turns as context, the LLM must identify which (if any) of the preceding turns the current one responds to.

**Stage 2 (segmentation):** Given a pair of speaker turns in which one responds to the other, the second stage aims to identify which exact part of the turn

[1] We use all-mpnet-base-v2: https://huggingface.co/sentence-transformers/all-mpnet-base-v2

[2] We use model versions `gpt-4o-2024-08-06` and `claude-3-5-sonnet-20241022` via their respective APIs.

responds to which exact part of the given preceding turn. Each (sub)part is called a segment.

**Stage 3 (classification):** The goal of the third stage is to classify each pair of responsive segments as mechanical or substantive.

Following best practices for LLM annotation tasks (Pangakis et al., 2023), we perform three runs of the first and third stages for each conversation. To maximize consistency, we drop any labels that appear in less than 2 out of the 3 runs. Full prompts can be found in Appendix D.

### 4.3 Human Annotation Task

In order to better understand how the models structure these conversations, we design and deploy a human annotation task to develop a human set of annotations and links to compare our computationally generated structures. We design the task to be completed on the crowdsourcing research platform Prolific and specify that crowdworkers need to be fluent in English.We pay each Prolific worker $17 an hour, $2 above minimum wage in our research institution's state. Each task is to annotate one full conversation, and the worker completion time ranged from 30 minutes to 1.5 hours. When first deployed, we invited workers to give feedback on the complexity of the task, and the feedback we received was that it was a complex task at first, but after a few turns, they were able to complete it with greater ease. The study was reviewed and approved by the Ethics Review Board.

We had three annotators per conversation, calculated the inter annotator agreement, and treated a majority vote as the human standard to which we compared the LLM and semantic similarity annotations. For the human annotations, we did not ask participants to distinguish between substantive and mechanical responsivity, and hope to evaluate this with human annotation in the future.

## 5 Evaluation

### 5.1 Dataset

To investigate responsivity as a metric for conversation quality and to evaluate our proposed methods, we use the Fora dataset (Schroeder et al., 2024), a dataset of 262 richly annotated facilitated conversations that were hosted with partner organizations seeking to engage their members and surface insights regarding issues like education, elections, and public health. Similar to the spirit of our work, the conversations in the Fora dataset center around

| | claude | claude-fs | gpt4o | gpt4o-fs | human-cons | sem_sim |
|---|---|---|---|---|---|---|
| claude | 1.00 | 0.81 | 0.76 | 0.62 | 0.67 | 0.31 |
| claude-fs | 0.81 | 1.00 | 0.74 | 0.68 | 0.57 | 0.32 |
| gpt4o | 0.76 | 0.74 | 1.00 | 0.71 | 0.57 | 0.32 |
| gpt4o-fs | 0.62 | 0.68 | 0.71 | 1.00 | 0.45 | 0.31 |
| human-cons | 0.67 | 0.57 | 0.57 | 0.45 | 1.00 | 0.22 |
| sem_sim | 0.31 | 0.32 | 0.32 | 0.31 | 0.22 | 1.00 |

Figure 2: Inter-annotator agreement matrix between annotation methods. Note: fs refers to few shot and cons refers to consolidated.

the sharing of stories and personal experiences, which has been shown (Kessler et al., 2024) to elicit higher levels of empathy and understanding (compared to the sharing of opinions, especially for polarized sides). The Fora conversations are mostly small-group conversations (median of 6 participants) that were held between 2019-2023, recorded with consent, and automatically transcribed. The organizer was able to redact and edit transcripts for any transcription mistakes following the conversation.

Alongside this corpus we include a set of conversation collections that are similarly small-group facilitated dialogues that are recorded and transcribed, but that we would expect to be structured differently. Specifically, we included recorded game-play of conversation-based games, a collection of facilitated deliberation sessions from a citizen's assembly developing policy recommendations, and recorded conversations of youth discussing themes in a youth-focused documentary. While these collections contain many similar core attributes, we anticipate them to be recognizably different based on our derived metric set.

### 5.2 Responsivity Annotation Evaluation

We evaluate differences between annotations using the Jaccard index. The Jaccard index, also known as the Jaccard similarity coefficient, quantifies the overlap of two sets as the size of their intersection divided by the size of their union. For our purposes, when two annotators provide identical annotations on a conversation turn, the Jaccard index will be 1, while a completely non-overlapping set of annotations yields a Jaccard index of 0.

Across the data, the average number of responses per turn annotated by GPT-4 across two conversations is 1.42. For Claude, 1.25. for humans, 1.04. Of those annotations, the average number of

substantive responses per speaker turn is 0.99 for GPT-4 and 0.78 for Claude. Within our annotated dataset, two conversations were a part of the Fora corpus's story and personal experience annotation scheme. Using those previously annotated conversation turns, we see that 29% of speaker turns are labeled as sharing a personal story/experience. For instances of personal stories and/or experiences, we observe a much higher responsivity rate to those speaker turns, and specifically a higher rate of *substantive* responses. Specifically, non-personal story contributions that were responded to received mechanical responses about 40% of the time, while story contributions that were responded to received mechanical responses about 30% of the time.

The inter-annotator agreement matrix in Figure 2 shows that Claude and GPT-4 were most aligned out of the annotation methods, followed by Claude and human annotation. The least aligned method of responsivity mapping is the semantic similarity approach, with no Jaccard index greater 0.32. Further, we see that inter-annotator agreement calculated through the Jaccard index for human annotation is low, with an average Jaccard index across all conversations of 0.592. One can see the inter-annotator agreement matrix for a single conversation visualized in Figure 3, showing alignment on par with the LLM to human alignment. Interestingly, the few shot LLM prompts generally yielded lower inter-annotator agreements than the one-shot methods. In Appendix C, we highlight moments of disagreement between human annotators and LLMs to exemplify how interpretations of responsivity depend on one's position, but the breadth of interpretations is reasonably bounded.

We then compared accuracy of LLM annotations against human annotations for substantive versus mechanical labels with a subset of 100 conversation snippets. Using the preceding context window of "possible responsivity links" for each turn, there was agreement between human and LLM annotators on 91.9% of the labels (10.6% links present for both, 81.3% absent for both). In cases where they disagreed, humans exclusively labeled a link 3.3% and LLMs 4.8% of the time. The analysis shows high levels of agreement, and reveals nuances between LLM and human understanding, such as LLMs label responses as substantive slightly more often than humans.

| | 0 | 1 | 2 |
|---|---|---|---|
| 0 | 1.000000 | 0.751964 | 0.711616 |
| 1 | 0.751964 | 1.000000 | 0.648597 |
| 2 | 0.711616 | 0.648597 | 1.000000 |

Figure 3: Human inter-annotator agreement matrix for conversation 1113.

## 6 Conversation Analysis

In the preceding sections, we proposed and evaluated approaches to responsivity annotation. While the semantic similarity based method had low agreement with human annotators, LLM-based methods performed well. A dialogue annotated with such responsivity links supports an examination of conversation structure, both visually (see Figure 1) and through derived summary metrics. We can see that links between speaker turns show how participants build upon one another, and fragmented sections of conversation show no interconnections. Side conversations are visibly distinct from main conversations in the flow, and highly impactful moments seem visible from their many interconnections.

In this section, we describe some conversation metrics that we show support characterization and differentiation of conversations. The first set of metrics derives directly from turn information, while the second set builds upon the responsivity annotations.

The first set includes simple measures such as the number of speakers, the total number of turns and total duration of the conversation, the number of facilitator turns and speaking time, and the corresponding percentages, and the variance in the number of turns across speakers. We also compute distributional features – namely, the Gini coefficient (Dorfman, 1979; Farris, 2010) of both the speaking time and number of turns to quantify how balanced (Gini coefficient near 0) or unbalanced (Gini coefficient near 1) these quantities are across speakers. Finally, we compute the conditional entropy on the speaker turn *sequence* to characterize the variability in speaker turn-taking. A perfectly consistent speaking order would yield a conditional entropy of 0, increasing to a maximum for a random speaker turn ordering.

Since a conversation is not simply a sequence of turns, but rather a sequence *connected* by participants responding to one another, we develop

additional metrics to characterize this responsivity structure. The simplest metrics are the rates of substantive and mechanical responsivity. However, since the preceding window used for responsivity annotations may include the speaker as well as the facilitator, we also calculate rates restricting to the subset of non-self and non-facilitator turns. As above, we calculate distributional metrics, but here quantify how actual *responses* are distributed across participants using the Gini coefficient. We compute the Gini coefficient on the distribution of substantive responsivity to quantify whether everyone was equally substantively responsive or whether it was concentrated on only a few participants (we also compute variations on the subset of preceding turns considered to exclude facilitator and self). Finally, we compute the conditional entropy of the distribution of who substantively responds to whom, helping quantify whether everyone generally responded to everyone else or whether participants were more selective in who they responded to. See Appendix A, Table 1 for a summary description of all features.

The conversation-level metrics support comparing between conversations and analysis of large conversation collections. For this evaluation, we analyzed conversations in the Fora Corpus as well as the youth documentary discussions (n=11), the citizen assembly (n=13), and game-play data (n=12). As described above, some of these collections have different purposes and formats.

To begin, we computed all features (23 in total) for each of the 101 conversations described in Appendix A. The features themselves are motivated by observations and experience with small-group dialogue. Since we know that some of these features are correlated (see Appendix B, Figure 7), so to support interpretability we identify groups of highly correlated features and take only a subset. We did this manually given the small number of features and our original goal of capturing certain aspects of conversation structure. For example, Figure 7 shows that `avg_subst_responded_rate` and a block of functionally related (but more specific) features are highly correlated. In this case, we decided to keep the base feature and the related "non-self" feature, since we felt it reflected an important distinction (preferring responses to others rather than oneself). We note that we did experiment with PCA on the original feature set, finding that only 9-10 features are needed to preserve 95% of the variance. However, the resultant components are much harder to interpret as they are linear combinations of the original 23 features.

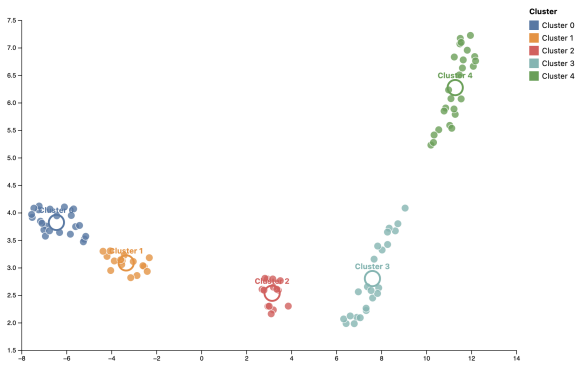


Figure 4: Cluster Map showing the 5 clusters.

This process yielded 12 features from the original 23, listed in Figure 5. These include both direct features (e.g. speaking time, percentages, etc.) as well as features derived from responsivity annotations (e.g. substantive and mechanical responsivity rates, entropy, etc.). We clustered the conversations using these features, first by applying UMAP (McInnes et al., 2018) to reduce the features to 3 dimensions, and then clustering with HDBscan (McInnes et al., 2017). This yielded 5 clusters, which we describe below. Further, we visualize the conversations in a 2-dimensional UMAP-reduced cluster-colored plot, in Figure 4.

## 7 Conversation Clusters Analysis Results

In the following section, we describe the clusters identified through our cluster analysis to accompany the centroids outlined in Figures 4 and 5.

**0: Facilitated, Dynamic Small Groups.** These conversations are shorter and involve smaller groups. They are marked by high equality in speaking time and high entropy in both turn-taking and responsivity, suggesting a free-flowing, dynamic exchange. Facilitators play a relatively prominent role in speaking, while substantive responsivity is at a moderate level.

**1: Participant-Driven, Substantively Engaged Dialogues.** This cluster has the highest rates of substantive responsivity, and the lowest rates of mechanical responsivity. With the lowest percentage of facilitator speaking time, these show participant driven, responsive conversations. This cluster holds mid-levels of entropy and gini coefficient, showing a balanced but structured flow.

**2: Structured, Unequal, Large-Group.** Conversations in this cluster are the largest and most un-

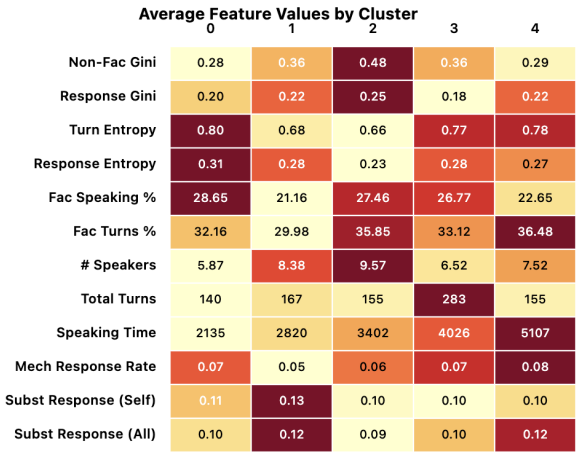


Figure 5: Heatmap showing the average feature values by cluster.

equal in terms of participation. They show the highest Gini coefficients and the lowest entropy, pointing to a highly structured and facilitator-dominated format. Substantive responsivity is the lowest, suggesting a less engaged or more top-down dynamic.

**3: High Turns, Disordered, Low Response.** These are long conversations with a high number of turns and slightly elevated entropy, implying less orderly interaction. Substantive responsivity is low, suggesting limited depth or follow-through in exchanges. Average speaking time time is modestly elevated, mechanical responsivity is higher and substantive is lower. That, tied with a slightly elevated facilitator speaking time suggests longer, disordered, less substantive, more facilitator driven conversation. This pattern often correlated with face-paced conversation games.

**4: Inclusive, High-Engagement, Long.** This cluster includes the longest conversations by speaking time and the highest percentage of facilitator turns though facilitator speaking time is quite low, suggesting they take many brief turns. Conversely, participant take few turns, but have the longest speaking time, suggesting long, extensive turns. It features the lowest speaking inequality, indicating a highly inclusive dynamic, and notably high substantive responsivity when self and facilitator contributions are excluded. The slightly elevated responsivity Gini suggests a less predictable, possibly more exploratory style of engagement.

## 8 Discussion

In this work, we argue that understanding constructive conversation requires first examining the underlying structure of contributions and responses. How participants listen, respond, and build upon one another is foundational to conversation, shaping its flow, emergent relationships, and overall quality. We build on prior work introducing responsivity as a key metric, applying it across conversation windows rather than as a summary statistic.

Expanding on the approach of Dowell et al. (2019), we prompt state-of-the-art LLMs to annotate conversations for responsivity and, for the first time, evaluate this method against human annotators. Our findings indicate that LLMs align more closely with human judgments than semantic similarity-based approaches. However, disagreement among human annotators highlights the inherent difficulty and subjectivity of the task. Notably, while LLMs do not perfectly match human annotations, inter-LLM-alignment is comparable to variations observed among humans

Using our approach, we further observe responsivity dynamics validating our expectations from the dialogue literature. For example, in facilitated discussions, participants are expected to be more responsive to one another than to facilitators—a pattern reflected in our responsivity data (Wilson and Prinzo, 2002; Schroeder et al., 2024). Likewise, sharing stories tends to foster stronger connections between participants than opinions, as signaled by higher responsivity, particularly substantive responsivity, to story-based contributions.

Building on these observations, we introduce a set of derived metrics designed to capture meaningful distinctions in conversation styles and structures. Applying these metrics to a diverse set of conversations, we demonstrate that clustering produces interpretable groupings that align with actual differences in conversational purpose, style, and structure. These findings suggest that our derived metrics effectively distinguish between conversation dynamics, further validating their utility in dialogue analysis.

The distinction between substantive and mechanical responsivity proves particularly valuable in characterizing conversation quality. Our analysis reveals that conversations with higher rates of substantive responsivity tend to exhibit different structural patterns than those dominated by mechanical responses. This finding supports theoretical frameworks from dialogue studies that emphasize meaningful engagement over mere acknowledgment as a marker of constructive interaction.

Furthermore, our conversation-level clustering

analysis demonstrates that responsivity patterns can meaningfully differentiate between conversation types. The five clusters we identified—ranging from "Facilitated, Dynamic Small Groups" to "Inclusive, High-Engagement, Long" conversations—each exhibit distinct responsivity signatures that align with their intended purposes and facilitation styles. This suggests that responsivity metrics capture not just individual interaction quality but also systemic properties of different conversational contexts.

### 8.1 Applications and Integration Opportunities

We see three primary implications of this work. First, conversation has long been and will continue to be a key medium for democratic civic participation. While extensive research examines non-constructive, toxic, and polarizing discourse, existing work on constructive communication has primarily focused on dyadic relationships (Gable et al., 2018; Rusbult et al., 1991), educational settings through collaborative learning research (Johnson and Johnson, 1999), and social-emotional learning frameworks (Weissberg et al., 2015). However, less attention has been given to systematically measuring and characterizing constructive communication patterns in multi-party civic dialogue contexts. Our work provides a method to support those designing constructive communication spaces by helping them assess and refine their interventions. This kind of reflection, aimed at improving conversational dynamics, builds on existing work such as Meeting Mediator and Keeper (Kim et al., 2008; Hughes and Roy, 2021; Adachi et al., 2015).

Second, while we do not apply our approach to online conversations in this work, we believe it holds significant potential for digital spaces. Comment sections of videos and news articles, subreddits, or threads on microblogging platforms could benefit from analyzing not only explicit replies but also responses that build upon previous contributions. This shift would allow us to examine both the structured conversational elements embedded in platform design and the more nuanced interactions experienced by participants—structures often invisible in traditional data analysis.

Finally, as more governance processes and public discourse move online through tools like Pol.is (Small et al., 2021) and Remesh (Konya et al., 2023), evaluating online conversations will become increasingly important. Digital communities require governance and moderation, and we believe this approach could support prosocial moderation by providing insights into the underlying conversational dynamics of online communities.

### 8.2 Future Work

While responsivity is a foundational metric, we do not see it as the sole indicator of constructive conversation. Drawing from both theory and practice, we recognize that elements such as personal story sharing, introducing new ideas, and other core aspects of dialogue also signal value and connection. We aim to expand our framework to incorporate a more comprehensive set of metrics in future work. With this expanded set of metrics, we also hope to build on Dowell's work by developing a taxonomy of participation types. If we cluster participants based on these characteristics, do clear patterns emerge? Finally, we seek to apply these metrics across "high" and "low quality" conversational types. Can this approach effectively analyze different conversation quality? Can it help identify constructive communication across varied contexts? And does it enhance our understanding of prosocial and constructive discourse in both digital and in-person spaces?

## 9 Conclusion

In this work, we map responsivity between conversation participants in multiple ways to reveal conversation structures. This work, we believe, can help lay the foundation for developing a more comprehensive set of conversation metrics so we may understand what makes conversations constructive and healthy. We compare human annotation, a semantic similarity approach, and a large language model approach to describe the alignment between these methods and various conversation structures, such as facilitator acts and storytelling. We then develop a set of metrics derived from the responsivity structure. Clustering and analysis on these derived metrics reveal qualitatively different kinds of conversations, which we find generally align with real differences in purpose and style of facilitated conversations. Looking forward, we believe this work will contribute meaningfully to conversation analysis in various domains within and beyond computational linguistics, ranging from the civic world, to computer-supported cooperative work, to online conversation spaces.

## 10 Limitations

There are several key limitations to this study. First, we have not systematically compared or evaluated the differences between the LLM, semantic similarity, and human annotations. While we understand to what degree they are aligned or not, we have not fully examined the points where they disagree. For example, are there certain kinds of turns that Claude appropriately annotates, that GPT and semantic similarity miss? We could discern this through further qualitative investigation into the disagreement points. Further, while we did collect human annotations, there was meaningful disagreement between human annotators. We are not certain why there was disagreement, on what kinds of speaker turns they disagreed, and if there were meaningful patterns across participants in their annotation styles. Finally, while LLMs show promising alignment with human annotations, they may still encode biases inherent in their training data. Future work could explore these potential biases within the conversation quality and dynamics context. We again hope to improve this study for future work by understanding more deeply these patterns through systematic, qualitative review. With respect to derived metrics, there are surely other metrics that can meaningfully characterize different aspects of conversation structure and dynamics. We have pursued this primarily in an unsupervised setting, but appropriate conversation labels could support determining which features are most salient or informative for the task.

## References


Hiroyuki Adachi, Seiko Myojin, and Nobutaka Shimada. 2015. Scoringtalk: a tablet system scoring and visualizing conversation for balancing of participation. In *SIGGRAPH Asia 2015 Mobile Graphics and Interactive Applications*, SA '15, New York, NY, USA. Association for Computing Machinery.

Danielle Allen. 2023. Justice by means of democracy. In *Justice by Means of Democracy*. University of Chicago Press.

Anthropic. 2024. Introducing Claude 3.5 Sonnet.

Shubham Atreja, Joshua Ashkinaze, Lingyao Li, Julia Mendelsohn, and Libby Hemphill. 2024. Prompt design matters for computational social science tasks but in unpredictable ways. *Preprint*, arXiv:2406.11980.

Mikhail Mikhaĭlovich Bakhtin. 2010. *The dialogic imagination: Four essays*, volume 1. University of texas Press.

Jiajun Bao, Junjie Wu, Yiming Zhang, Eshwar Chandrasekharan, and David Jurgens. 2021. Conversations gone alright: Quantifying and predicting prosocial outcomes in online conversations. In *Proceedings of the Web Conference 2021*, WWW '21, page 1134–1145, New York, NY, USA. Association for Computing Machinery.

Jake Barrett, Kobi Gal, Loizos Michael, and Dan Vilenchik. 2024. Beyond the echo chamber: modelling open-mindedness in citizens' assemblies. *Autonomous Agents and Multi-Agent Systems*, 38(2):30.

David A. Broniatowski. 2012. Extracting social values and group identities from social media text data. In *2012 IEEE 14th International Workshop on Multimedia Signal Processing (MMSP)*, pages 232–237.

Youngjin Chae and Thomas Davidson. 2023. Large language models for text classification: From zero-shot learning to instruction-tuning.

Kent K Chang, Danica Chen, and David Bamman. 2023. Dramatic conversation disentanglement. *arXiv preprint arXiv:2305.16648*.

Daejin Choi, Jinyoung Han, Taejoong Chung, Yong-Yeol Ahn, Byung-Gon Chun, and Ted Taekyoung Kwon. 2015. Characterizing conversation patterns in reddit: From the perspectives of content properties and user participation behaviors. In *Proceedings of the 2015 acm on conference on online social networks*, pages 233–243.

Claudia Chwalisz. 2019. A new wave of deliberative democracy. *Carnegie Europe*, 26:1–6.

Claudia Chwalisz. 2020. *Reimagining democratic institutions: Why and how to embed public deliberation*. OECD.

Bosheng Ding, Chengwei Qin, Linlin Liu, Yew Ken Chia, Boyang Li, Shafiq Joty, and Lidong Bing. 2023. Is GPT-3 a good data annotator? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11173–11195, Toronto, Canada. Association for Computational Linguistics.

Robert Dorfman. 1979. A formula for the gini coefficient. *The Review of Economics and Statistics*, 61(1):146–149.

Nia MM Dowell, Tristan M Nixon, and Arthur C Graesser. 2019. Group communication analysis: A computational linguistics approach for detecting sociocognitive roles in multiparty interactions. *Behavior research methods*, 51:1007–1041.

Ritam Dutt, Zhen Wu, Kelly Shi, Divyanshu Sheth, Prakhar Gupta, and Carolyn Penstein Rose. 2024. Leveraging machine-generated rationales to facilitate social meaning detection in conversations. *Preprint*, arXiv:2406.19545.

Selen A Ercan, Hans Asenbaum, Nicole Curato, and Ricardo F Mendonça. 2022. *Research methods in deliberative democracy*. Oxford University Press.

Frank A. Farris. 2010. The gini index and measures of inequality. *The American Mathematical Monthly*, 117(10):pp. 851–864.

Shelly L Gable, Harry T Reis, Emily A Impett, and Evan R Asher. 2018. What do you do when things go right? the intrapersonal and interpersonal benefits of sharing positive events. In *Relationships, well-being and behaviour*, pages 144–182. Routledge.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. Chatgpt outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30):e2305016120.

Xingwei He, Zhenghao Lin, Yeyun Gong, A-Long Jin, Hang Zhang, Chen Lin, Jian Jiao, Siu Ming Yiu, Nan Duan, and Weizhu Chen. 2024. AnnoLLM: Making large language models to be better crowdsourced annotators. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 6: Industry Track)*, pages 165–190, Mexico City, Mexico. Association for Computational Linguistics.

Fan Huang, Haewoon Kwak, and Jisun An. 2023. Is chatgpt better than human annotators? potential and limitations of chatgpt in explaining implicit hate speech. In *Companion Proceedings of the ACM Web Conference 2023*, WWW '23 Companion, page 294–297, New York, NY, USA. Association for Computing Machinery.

Margaret A Hughes and Deb Roy. 2021. Keeper: A synchronous online conversation environment informed by in-person facilitation practices. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–14.

Judith E. Innes and David E. Booher. 2004. Reframing public participation: strategies for the 21st century. *Planning Theory & Practice*, 5(4):419–436.

William Isaacs. 1999. *Dialogue: The art of thinking together*. Crown Currency.

David W Johnson and Roger T Johnson. 1999. Making cooperative learning work. *Theory into practice*, 38(2):67–73.

Dan Jurafsky, Rajesh Ranganath, and Dan McFarland. 2009. Extracting social meaning: Identifying interactional style in spoken conversation. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 638–646, Boulder, Colorado. Association for Computational Linguistics.

Daniel Kessler, Dimitra Dimitrakopoulou, and Deb Roy. 2024. Hearing Personal Experiences Improves Social Evaluations Compared to Personal Opinions, Especially for Polarized Parties. *SSRN preprint*. [Accessed 09-10-2024].

Taemie Kim, Agnes Chang, Lindsey Holland, and Alex Sandy Pentland. 2008. Meeting mediator: enhancing group collaborationusing sociometric feedback. In *Proceedings of the 2008 ACM conference on Computer supported cooperative work*, pages 457–466.

Andrew Konya, Lisa Schirch, Colin Irwin, and Aviv Ovadya. 2023. Democratic Policy Development using Collective Dialogues and AI. *arXiv preprint*. ArXiv:2311.02242.

Katerina Korre, Dimitris Tsirmpas, Nikos Gkoumas, Emma Cabalé, Dionysis Kontarinis, Danai Myrtzani, Theodoros Evgeniou, Ion Androutsopoulos, and John Pavlopoulos. 2025. Evaluation and facilitation of online discussions in the llm era: A survey. *arXiv preprint arXiv:2503.01513*.

Jonathan K. Kummerfeld, Sai R. Gouravajhala, Joseph J. Peper, Vignesh Athreya, Chulaka Gunasekara, Jatin Ganhotra, Siva Sankalp Patel, Lazaros C Polymenakos, and Walter Lasecki. 2019. A large-scale corpus for conversation disentanglement. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3846–3856, Florence, Italy. Association for Computational Linguistics.

Alyssa Lees, Vinh Q. Tran, Yi Tay, Jeffrey Sorensen, Jai Gupta, Donald Metzler, and Lucy Vasserman. 2022. A new generation of perspective api: Efficient multilingual character-level transformers. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '22, page 3197–3207, New York, NY, USA. Association for Computing Machinery.

Zhuoyan Li, Hangxiao Zhu, Zhuoran Lu, and Ming Yin. 2023. Synthetic data generation with large language models for text classification: Potential and limitations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10443–10461, Singapore. Association for Computational Linguistics.

Leland McInnes, John Healy, and Steve Astels. 2017. hdbscan: Hierarchical density based clustering. *The Journal of Open Source Software*, 2(11).

Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. 2018. Umap: Uniform manifold approximation and projection. *Journal of Open Source Software*, 3(29):861.

OpenAI. 2024. GPT-4o System Card.

Nicholas Pangakis, Samuel Wolken, and Neil Fasching. 2023. Automated Annotation with Generative AI Requires Validation. *arXiv preprint*. ArXiv:2306.00176.

Mattes Ruckdeschel. 2025. Just read the codebook! make use of quality codebooks in zero-shot classification of multilabel frame datasets. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6317–6337, Abu Dhabi, UAE. Association for Computational Linguistics.

Caryl E Rusbult, Julie Verette, Gregory A Whitney, Linda F Slovik, and Isaac Lipkus. 1991. Accommodation processes in close relationships: Theory and preliminary empirical evidence. *Journal of Personality and social Psychology*, 60(1):53.

Martin Saveski, Brandon Roy, and Deb Roy. 2021. The structure of toxic conversations on twitter. In *Proceedings of the Web Conference 2021*, WWW '21, page 1086–1097, New York, NY, USA. Association for Computing Machinery.

Emanuel A Schegloff. 2007. *Sequence organization in interaction: A primer in conversation analysis I*, volume 1. Cambridge university press.

Hope Schroeder, Deb Roy, and Jad Kabbara. 2024. Fora: A corpus and framework for the study of facilitated dialogue. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13985–14001, Bangkok, Thailand. Association for Computational Linguistics.

Christopher Small, Michael Bjorkegren, Timo Erkkilä, Lynette Shaw, and Colin Megill. 2021. Polis: Scaling Deliberation by Mapping High Dimensional Opinion Spaces. *RECERCA. Revista de Pensament i Anàlisi*, 26(2). Publisher: Universitat Jaume I.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. Mpnet: Masked and permuted pre-training for language understanding. *Advances in neural information processing systems*, 33:16857–16867.

Karen Tracy and Margaret Durfy. 2007. Speaking out in public: citizen participation in contentious school board meetings. *Discourse & Communication*, 1(2):223–249.

Shuohang Wang, Yang Liu, Yichong Xu, Chenguang Zhu, and Michael Zeng. 2021. Want to reduce labeling cost? GPT-3 can help. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4195–4205, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Roger P Weissberg, Joseph A Durlak, Celene E Domitrovich, and Thomas P Gullotta. 2015. Social and emotional learning: Past, present, and future.

Robin J Wilson and Michelle Prinzo. 2002. Circles of support: A restorative justice initiative. *Journal of Psychology & Human Sexuality*, 13(3-4):59–77.

Ziang Xiao, Xingdi Yuan, Q. Vera Liao, Rania Abdelghani, and Pierre-Yves Oudeyer. 2023. Supporting qualitative analysis with large language models: Combining codebook with gpt-3 for deductive coding. In *Companion Proceedings of the 28th International Conference on Intelligent User Interfaces*, IUI '23 Companion, page 75–78, New York, NY, USA. Association for Computing Machinery.

Lexin Zhou, Youmna Farag, and Andreas Vlachos. 2024. An llm feature-based framework for dialogue constructiveness assessment. *arXiv preprint arXiv:2406.14760*.

Rongxin Zhu, Jey Han Lau, and Jianzhong Qi. 2021. Findings on conversation disentanglement. In *Proceedings of the 19th Annual Workshop of the Australasian Language Technology Association*, pages 1–11, Online. Australasian Language Technology Association.

Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2024. Can large language models transform computational social science? *Computational Linguistics*, 50(1):237–291.

## A Full List of Features with Definitions

| Feature | Definition |
|---|---|
| speaking_time_gini_coefficient | Measures inequality in total speaking time across all participants. Higher values indicate more unequal participation. |
| turn_distribution_gini_coefficient | Measures inequality in the number of turns taken by each participant. |
| non_facilitator_speaking_gini_coefficient* | Speaking time inequality among non-facilitator participants only. |
| non_facilitator_turn_gini_coefficient | Turn-taking inequality among non-facilitator participants. |
| gini_subst_responded_rate_nonself* | Inequality in substantive response rates, excluding self-responses. |
| gini_subst_responded_rate_nonself_nonfac | Substantive response rate inequality, excluding self- and facilitator-directed responses. |
| gini_subst_responded_rate_nonself_exclfac | Substantive response rate inequality, excluding responses from the facilitator. |
| gini_subst_responded_rate_nonself_nonfac_exclfac | Most restrictive: excludes responses to self and facilitator, and includes only non-facilitator targets. |
| turn_sequence_entropy* | Entropy of the speaker turn sequence. Higher values suggest less predictable (more disordered) turn-taking. |
| substantive_responsivity_entropy* | Entropy of how substantive responses are distributed across speakers. |
| facilitator_speaking_percentage* | Percentage of total speaking time contributed by the facilitator. |
| facilitator_turns_percentage* | Percentage of turns taken by the facilitator. |
| num_turns_facilitator | Raw count of turns taken by the facilitator. |
| num_observed_speakers* | Number of unique speakers in the conversation. |
| total_turns_in_conversation | Total number of turns across all speakers. |
| total_speaking_time_seconds* | Total amount of speaking time in seconds. |
| turn_count_variance | Variance of distribution of number of turns across participants. |
| avg_subst_responded_rate | Average rate at which participants give substantive responses. |
| avg_mech_responded_rate* | Average rate at which participants give mechanical responses. |
| avg_subst_responded_rate_nonself* | Substantive response rate directed at others (excluding self-responses). |
| avg_subst_responded_rate_nonfac | Substantive responses excluding facilitator as response reciever. |
| avg_subst_responded_rate_nonself_exclfac | Substantive responses excluding self and directed at others, excluding the facilitator. |
| avg_subst_responded_rate_nonself_nonfac_exclfac* | Most restrictive substantive response average — excludes both self and facilitator, as responder or target. |

Table 1: Definitions of conversation analysis features. Features used in the final, reduced set have an astrisk.

## B Initial Clusters

In an earlier version of conversation clustering, we used the full set of 23 features. We applied UMAP (McInnes et al., 2018) to these features, reducing to 5 dimensions, followed by clustering with HDBscan (McInnes et al., 2017). For visual inspection, we also applied UMAP to the full feature set to obtain a 2-dimensional visualization, shown in Figure 6. The characteristics (i.e. average feature values) are provided in Table 2. We named the clusters based on their feature characteristics, and describe the clusters below.

| Feature | Cluster 0 | Cluster 1 | Cluster 2 | Cluster 3 | Cluster 4 | Cluster 5 | Cluster 6 |
|---|---|---|---|---|---|---|---|
| Group Size | 23 | 18 | 14 | 11 | 12 | 15 | 7 |
| Speaking Time Gini | 0.303 | 0.412 | 0.297 | 0.452 | 0.388 | 0.370 | 0.329 |
| Turn Sequence Entropy | 0.775 | 0.780 | 0.803 | 0.653 | 0.743 | 0.697 | 0.781 |
| Substantive Responsivity Entropy | 0.277 | 0.257 | 0.310 | 0.254 | 0.278 | 0.304 | 0.321 |
| Facilitator Speaking % | 22.293 | 31.070 | 28.356 | 24.806 | 22.963 | 19.467 | 33.684 |
| Avg Subst Responded Rate | 0.122* | 0.075 | 0.143* | 0.101 | 0.106 | 0.116 | 0.071 |
| Total Turns | 137.696 | 426.444* | 103.500 | 113.818 | 107.417 | 127.267 | 227.286 |

Table 2: Clusters from initial clustering on all features, with average feature values for selecte features. Asterisks (*) indicate values deviating more than 1 standard deviation from the global mean.

**Dialogue Cluster:** Cluster 0 has medium to low Gini levels for speaking time, slightly elevated turn sequence entropy and facilitator speaking turns percentage, and about average Gini coefficients for responsivity. This cluster contains the majority of the Fora corpus, facilitated dialogues where the facilitator holds the space, takes many turns, and ensures a balance of speaking opportunities. Further, most of these conversations are on Zoom, consistent with increased facilitator speaking turns percentage.

Cluster 4 exhibits some similar characteristics, though with higher speaking time and turn distribution Gini coefficients. This holds the second greatest number of Fora conversations and repeated facilitators distributed across clusters 0 and 4. For these reasons, we define this cluster set as being the most dialogue-oriented clusters.

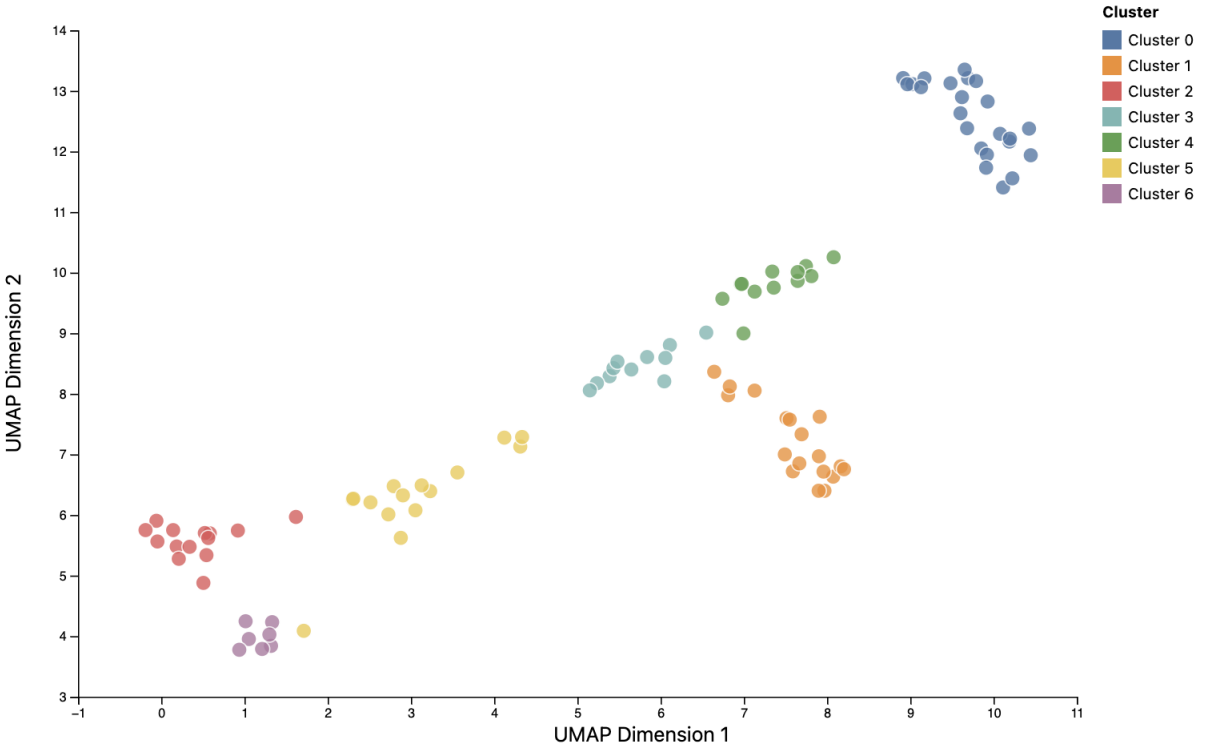


Figure 6: Cluster Map showing the 7 clusters.

**Dynamic Games:** Cluster 1 has very low substantive responsivity rates, medium levels of responsivity Gini coefficients, and medium to high speaking time Gini coefficients (both with and excluding facilitator turns), with by far the highest total turn counts. We see this cluster contains all game-based, in-person conversations with rapid turn-taking, unequal speaking time, and low responsivity. The games are split perfectly across this cluster and Cluster 6. Cluster 6 has the lowest responsivity Gini rates, the lowest average substantive or mechanical responsivity, and the highest turn-sequence and responsivity entropy, as well as the highest percentage of facilitator speaking time. The game played in this cluster is meant to mimic dialogue and has more structured turn-taking and higher responsivity entropy, showing the few but significant distinctions between the clusters.

**Responsive and Balanced:** Cluster 2 has the lowest turn and speaking time Gini coefficients, as well as low average Gini responsivity coefficients. It does exhibit high overall responsivity and the highest turn-sequence entropy. This suggests a highly balanced and responsive conversation, with a well mixed turn ordering between participants. This cluster contains second most conversations from the student assembly and conversations from the youth documentary and could be described as natural feeling and rich.

**Unequal and Predictable:** Cluster 3 has the highest Gini coefficients with the most unequal speaking time and rates of responsivity, along with the lowest turn sequence and responsivity entropy, suggesting they are highly unequal and very regimented speaking order. On average, they have middle rates of responsivity overall. This cluster contains one facilitator from the youth project and two groups of conversations with one or two individuals who often spoke extensively without the others contributing to the same degree. Based on these characteristics, we describe this cluster as one with unequal contributions, but a highly predictable structure.

**Deliberation and Discussion:** Cluster 5 has the highest responsivity rates, the lowest rates of facilitator speaking and turn-taking percentages, and average speaking time and responsivity Gini rates. This cluster contained almost all the conversations with youth discussing the documentary and the student assembly conversations. Their characteristics suggest debate, deliberation, and low facilitator intervention.

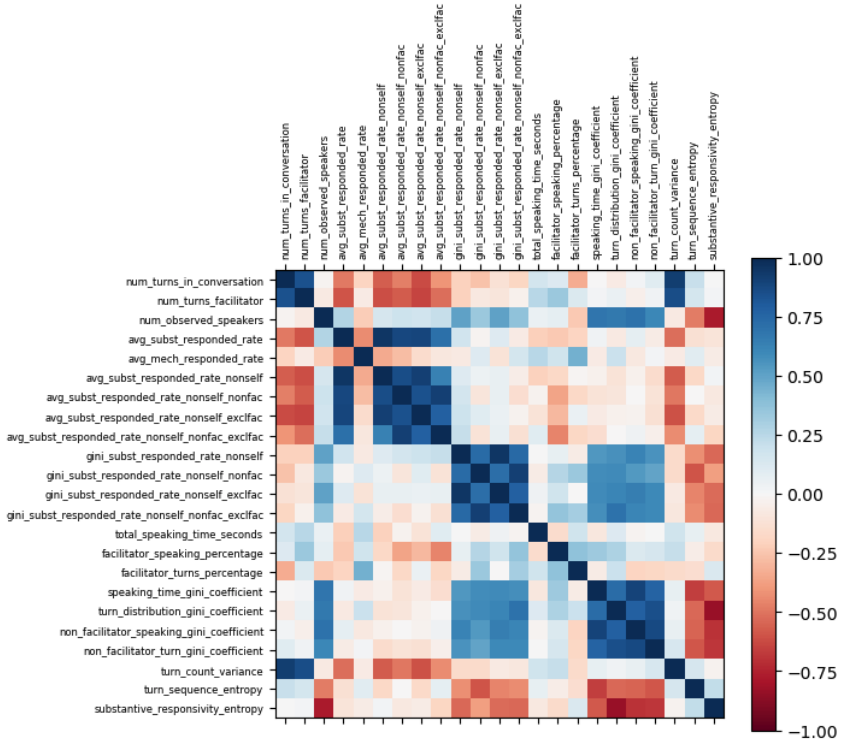


Figure 7: Heatmap showing the correlation between all 23 features. Note the block structure of both positively and negatively correlated features, primarily corresponding to variations on a fundamental measurement (e.g. variations on the speaking time and turn distribution Gini coefficient.)

## C Inter-Annotator Disagreement Examples

To unpack inter-annotator agreement scores, consider this excerpt from conversation 1093 (Table 3):

Consider turns 20-29 – GPT4o and Claude labels turn 29 as a response to turn 20, while semantic similarity labels it as responsive to turn 27, and the human annotator "majority vote" labels it a response to turn 28, with an overall net effect of reducing the annotation agreement (Jaccard) scores between human and machine. In fact, the 6 individual human annotators (note: most conversations annotated by 3 annotators) labeled this turn as follows: a0: [28], a1: [20, 28], a2: [28], a3: [28], a4: [20, 24, 28], a5: [27, 28], netting as 6 votes for turn 28, 2 votes for turn 20, and 1 vote for turns 24 and 27. The somewhat conservative "majority vote" strategy that we used takes only those labels submitted by at least half of the annotators, yielding only label 28 for this turn.

In reading the transcript, it is clear that turn 29 is a response to both the preceding turn (turn 28) and the question by the facilitator (turn 20), as provided by the LLMs. First, in turn 28 Justyce explicitly asks Maggie to share, which she does in turn 29. But in fact, the content of Maggie's turn (turn 29) is addressing the prompt by Justyce in turn 20. We would argue that both of these are valid, and perhaps best would be to label it as both.

An important point to make regarding the individual human annotations for turn 29 above is that they are actually quite consistent and specific – they do not span all possible preceding turns, but are concentrated on only a few. Having multiple humans randomly labeling would have produced lower inter-annotator agreement than what is shown in Figure 3. However, in future work we may wish to introduce additional mechanisms to ensure annotators spend sufficient time reviewing their selections.

We highlight another example in Table 4. In this instance, 3 humans annotators annotated speaker turn 22 to be responsive to the following turns: 16, [16, 19], [19, 20]. GPT annotated it as responsive to the following turns [19, 20]. This case is slightly more blurry. For example, turn 16 is where the facilitator, Fiona, prompts the group with a question. While the first two annotators perceive Kristel's turn as responding directly to that initial prompt, two annotators perceive her as building upon and responding

| Turn | Speaker | Utterance |
|---|---|---|
| 20 | Justyce | I will also say mine just so you can get a chance to know who I am. My name is Justyce, I am a student at UW Madison, and I'm a senior and I'm a local voices network intern. My pronouns are she, her, hers. And a value that is important to me is respect. I feel like we learned from a very young age that you should treat others the way you'd like to be treated. So that is a big aspect. Now, I'd like to invite you to share a little about your background. Take a minute and think about a personal story from your life that has shaped who you are. This could be from your childhood, adulthood, or work life, or something you saw happen to someone else. And we can go backwards this time. So Elizabeth, would you like to go first? |
| 21 | Elizabeth | Can I have another minute to think? |
| 22 | Justyce | Yeah. Meghan, are you ready at all? |
| 23 | Meghan | By life event, do you mean a memory that we hold closely to ourselves or...? |
| 24 | Justyce | Yeah, that works too. Just something about your background. It doesn't have to be anything specific. And if it's a memory, that works too. |
| 25 | Meghan | Okay. I guess I'm interested in diversity, equity, and inclusion because as a woman in STEM, it's hard to be heard compared to other people. So I value intersectional environmentalism, and also recognize that microaggressions are still prevalent in today's society. |
| 26 | Justyce | All right, thank you. And Amelia, are you able to go? |
| 27 | Amelia | Yeah. So I guess what I wanted to share was that I'm adopted from China. My parents are Caucasian, but I'm Asian. That played a big role in my upbringing—trying to be comfortable with that. It's kind of weird seeing my family in public, since we don't all look the same. That taught me to be more accepting of other people. |
| 28 | Justyce | Yeah. I definitely know how you feel there. Well, not the adoption part, but being mixed race and along those lines. And Maggie, would you like to go? |
| 29 | Maggie | Yeah. I grew up Catholic, which I know is one of the major religions in our society, but the town I grew up in was pretty Jewish. I don't know the exact proportion, but I... |

Table 3: Transcript excerpt of participants sharing background stories and values.

to herself and Libby who spoke right after. Again, no answer is objectively wrong. Turn 22 does seem to build upon the initial contribution from Kristel, and then further on Libby's comments, and may not obviously respond directly to the initial prompt itself, but it does add greater detail to the answer to the prompt and technically responds to it. We hope this example further highlights both the ambiguity of the task, but also the breadth of interpretations. While there are multiple interpretations, none are very far from one another, and out of 10 possible links, simple disagreements emerge around one or two of those links. Again, we show that annotations are quite consistent, but difference can emerge around the nuances of responsivity.

| Turn | Speaker | Utterance |
|---|---|---|
| 16 | Fiona | Perfect, thank you. Our second question is if you could share in a few words what you feel makes a thriving and resilient community? |
| 17 | Kristel | That's hard to put in a few words. |
| 18 | Fiona | You can use more than a few words. We have a couple people here. So you have, there's plenty of time. |
| 19 | Kristel | There's so many angles to approach that from. I could argue that communication is a huge part of that. I think access to resources is our absolute bottom line for that. Access to resources so that we're not living in food insecurity—especially with Maine and Kennebec County specifically having such high levels of food insecurity—that impacts every other area of someone's life. It's not just about feeling hungry. That's going to impact our young folks' ability to engage in education, which then impacts their future, physical development, and brain development. So something as seemingly simple as food is a building block to getting that thriving, resilient community. |
| 20 | Libby | Yeah. And I would just add... Excuse me, I apologize. Something I actually learned from Kristel, who is a great mentor for me. We all play a piece in this puzzle of creating a thriving, resilient community. I think a big piece of that is support, empathy, and understanding that we are all human and doing our best to better ourselves and the environment for others. A big part of our work lately has focused on recognizing strengths and the resources around us, but also understanding that we're all trying to make it through. Empathy and support are big pieces, along with addressing food insecurity, and making sure people are connected to shelter and food. It's hard to maintain mental health without those needs met. |
| 21 | Fiona | Great. |
| 22 | Kristel | I tend to go straight into something more tangible, given my background as a social worker. I want to see those pieces meeting a baseline for our folks and then obviously move into higher needs. Just wanted to give context as to why the mental health person went with food. |

Table 4: Transcript excerpt on defining a thriving and resilient community.

# D LLM Prompts

## D.1 Stage 1 (turn-level linking)

**System Instructions:**

```
Your task is to draw connections between the current, most recent conversation turn and the
preceding speaker turns in terms of **Responsivity**: the tendency of an individual to respond
(or not) to the contributions of their collaborative peers.

Now, you will be provided with an excerpt of a conversation, indexed by speaker turn id. Your
response should be in JSON according to the format specified below.
```

**Prompt Instructions:**

```
**Conversation excerpt:**
{excerpt}

**Current turn**
{current}

**Output instructions:**
Step 1: Consider the above conversation excerpt.
Step 2: Consider the current turn and whether it responds to any preceding turn.
Step 3: If it does, identify the preceding turn id(s) it specifically responds to in the
"link_turn_id" field. For not responsive segments, mark ["NA"] in the "link_turn_id" field.

Respond in JSON as follows:

{{
"link_turn_id": List<id of turn(s) responding to if applicable, otherwise ["NA"]>
}}
```

## D.2 Stage 2 (segmentation)

**System Instructions:**

```
Your task is to draw connections between two speaker turns in a conversation. Given two speaker
    turns in which one directly responds to the other, your task is to identify what specific
    part(s) of the second turn responds to what specific part(s) of the first.

Your response should be in JSON according to the format specified below.
```

**Prompt Instructions:**

```
**Speaker Turn 1:**
{speaker_turn_1}

**Speaker Turn 2:**
{speaker_turn_2}

**Output instructions:**
Step 1: Consider the above, in which {speaker_2} responds to {speaker_1}.
Step 2: Identify the part of Speaker Turn 2 that specifically responds to something in the
    previous turn. This should be an exact quote from Speaker Turn 2.
Step 3: Identify the part of Speaker Turn 1 that the above is directly responding to. This
    should be
an exact quote from Speaker Turn 1.

Respond in JSON as follows:

{{
"step_2": Str<your response to step 2>,
```

```
"step_3": Str<your response to step 3>
}}
```

### D.3 Stage 3 (classification)

**System Instructions:**

```
Your task is to draw connections between two speaker turns in a conversation that respond to
    each other. Each responsive speaker turn can be either **substantive** or **mechanical**:

- Substantive Responsivity refers to an interaction where one person meaningfully engages with
    what another has said. It captures how much a speaker reflects back, builds upon, inquires
    about, or connects to other ideas, emotions, or experiences shared by the previous speaker,
    or answers a meaningful question from a previous speaker.

- Mechanical Responsivity, on the other hand, occurs when a speaker responds in a way that
    acknowledges or moves the conversation forward but does not add substantial new content.
    These responses may include polite phrases, conversational hand-offs, or social cues.

Your response should be in JSON according to the format specified below.
```

**Prompt Instructions:**

```
**Speaker Turn 1:**
{speaker_turn_1}

**Speaker Turn 2:**
{speaker_turn_2}

**Output instructions:**
Step 1: Consider the above, in which {speaker_2} responds to {speaker_1}.
Step 2: Determine whether Speaker Turn 2 responds mechanically OR substantively to Speaker Turn
    1.
Step 3: If it truly has elements of both mechanical and substantive responsivity, then it should
     be considered substantive.

Respond in JSON as follows:

{{
"label": Str<"responsive_mechanical", or "responsive_substantive">,
}}
```

## E Annotation Task

Please read the following conversation turns:

**1**Opal: So, I've hit record and y'all should've seen that consent when we first came in as well, that we're going to record this. Any questions around that?

**2**Hannah: No.

**3**Opal: Awesome. Awesome. So what I'll do is, when we answer the different questions, I'll just go in the same order. So I'm just looking at my screen, so I'll go to Hannah and then Patrick and then Keetra, if you want to get in, fine. If not, no worries there as well. So that'll just be our order and you have the right at any of these if you just want to pass, don't feel like answering it, you can do that too and that's absolutely fine as well. It's a small group, so it's nice. Again, we can take our time with any of these as well. But I'll go ahead and just also say that my name is [Opal]. My pronouns are she, her and hers. And the first question is, what value or trait is important to you? So a value or trait that's important to me is compassion for others but as well as myself. So Hannah, do you want to introduce yourself a little bit and then say what trait is important to you, if you have one?

**4**Hannah: Yeah. So my name is Hannah. Pronouns are she, her, hers. And I would say a value that I find really important is honesty and parallel to that would be just transparency too. I think having that value, implementing that in your personal, professional, close, far away, all types of relationships is just really important to me.

**5**Opal: Absolutely.

**6**Patrick: Hey everybody, I'm Patrick [redacted]

**7**Opal: Hey, Patrick.

**8**Patrick: It's [redacted] like Cake. I'm in Waverly, Iowa. He, him, his, I guess, how you say it?

**9**Opal: Yap.

**********

**10**Patrick: A value or trait? Gosh. Opal, I'm going to pause. I'm not sure. These are going to be hard questions, because there's just one answer you could give, right?

**********

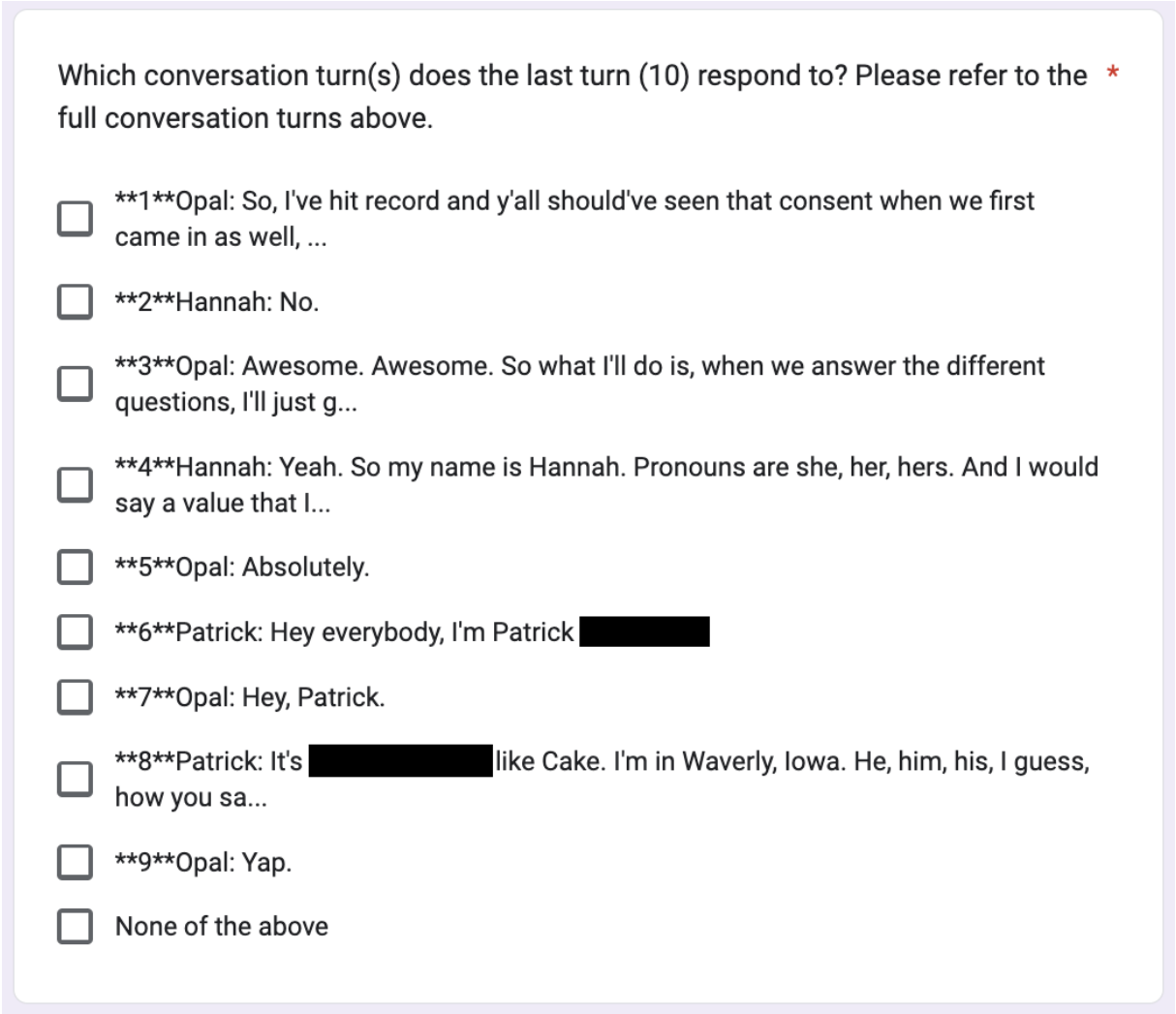


Figure 8: Form showing the human annotation task.