The WHoW Framework: An Automatic Approach to Cross-Scenario Conversational Moderation Analysis through Why, How, and Who

Anonymous ACL submission

Abstract

This study proposes the WHoW framework, an automatic approach for analyzing the facilitation strategies of conversational moderation across different scenarios. The framework breaks down moderation decision-making into three key components: motives, dialogue acts, and target speaker. Using this framework, we annotated 5,657 moderation sentences with human input and 15,494 sentences using GPT-4o across 196 episodes from two settings: the Intelligent Squared TV Debate and the RoundTable Radio Panel Discussion. Comparative analysis of these settings demonstrates the framework's cross-domain generalisability, revealing distinct moderation strategies: debate moderators emphasize coordination and facilitate interaction through questions and instructions, while panel discussion moderators prioritize information provision and actively participate in discussions, but are less involved in fostering inter-speaker interactions. This framework shows potential as a tool for exploring and comparing a broader range of moderation scenarios and could be expanded for the development of moderator agents. [1]

1 Introduction

Conversational moderation typically involves a moderator who upholds an impartial stance and interest, to facilitate and coordinate discussions among participants through conversation (Wright, 2009). Moderation occurs in diverse human interactive settings, however, the role of the moderator varies from hosts of debates (Thale, 1989; Zhang et al., 2016), judges in judicial processes (Danescu-Niculescu-Mizil et al., 2012), to therapists in group therapy sessions (Jacobs et al., 1998).

While there are various definitions of moderation across different domains (Grimmelmann, 2015; Vecchi et al., 2021; Friess and Eilders, 2015; Trénel, 2009) the concept is generally characterized as a form of discourse optimization mechanism with the essential objectives of: (1) mitigating: preventing and policing negative behaviors, such as personal attacks; (2) facilitating: promoting positive and constructive results, such as knowledge generation and consensus building; and (3) participating: ensuring balance and open participation opportunities for all members.

Extensive research has focused on content moderation analysis and automation in online spaces, primarily aimed at mitigating negative behaviors and intervening through actions such as post deletion (Gorwa et al., 2020; Park et al., 2021; Wulczyn et al., 2017; Falk et al., 2024). However, there has been relatively little exploration into how moderators facilitate positive outcomes and balance partic-

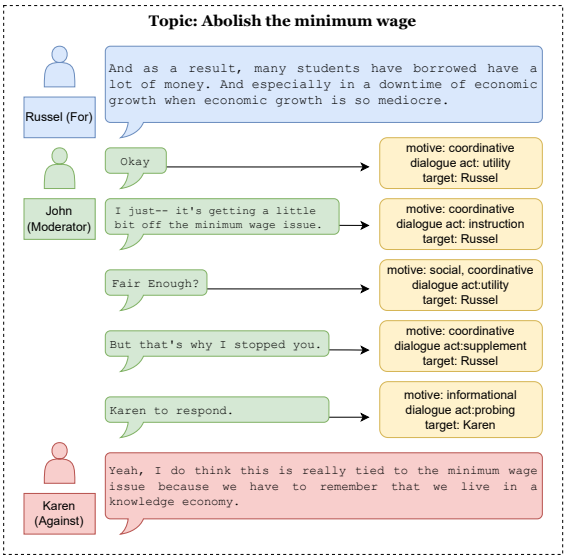


Figure 1: Example of a moderated conversation and annotation using the WHoW framework. Blue, green, and red colors represent the supporting team, moderator, and opposing team in the debate, respectively. The peach-colored boxes contain the annotations for the corresponding moderator sentences.

[1] Our code, dataset can will be released at Github after the anonymous period.

ipation through conversational engagement.

This study seeks to understand human conversational moderation practices by analyzing the decision-making processes of human moderators. It specifically focuses on how moderators facilitate discussions under various motives and balance participation among speakers. Additionally, the study aims to develop an automated tool capable of analyzing moderators' strategies on a large scale across diverse domains. Our tool and findings can support the development of moderator training or assessment, and potentially inform development of automated moderator agents such as in online discussions where human moderation does not scale.

We developed the WHoW analytical framework that breaks down the moderation decision-making process into three key components: motives (Why), dialogue acts (How), and target speaker (Who). Using this framework, we analyzed transcripts of human conversational moderation in two distinct contexts: the Intelligent Squared (INSQ) TV Debate Corpus and the Roundtable Radio Panel (RTRP). We began by annotating the test and development sets with human annotations, which were then used to create and evaluate prompts for Large Language Models (LLMs). These prompts were subsequently employed to automatically annotate a larger set. This automated annotation pipeline allowed us to analyze and compare the strategies to facilitate and balance participation used in the two scenarios.

Our key contributions are:

1. We developed an automatic analytic framework, powered by GPT-4o(OpenAI, 2024), that characterizes conversational moderation across different scenarios using three dimensions: motives (Why), dialogue acts (How), and target speaker (Who), effectively capturing the complexity of the task.

2. Utilizing the framework, we have annotated a large dataset of moderated multi-party conversations, encompassing two distinct scenarios: debates and panel discussions. This dataset comprises a total of 5,657 human-annotated sentences and 15,494 sentences annotated using GPT-4o.

3. By analyzing two conversational settings—debates and panel discussions—we demonstrate the framework's cross-domain generalizability, uncovering distinct moderation strategies. Debate moderators focus on coordination and facilitate interactions through questions and instructions, while panel discussion moderators prioritize information delivery and actively engage in discussions, but are less involved in promoting inter-speaker interactions.

2 Related Work

Conversational moderation is a complex task that requires consideration of multiple dimensions when making intervention decisions. This task takes place in multi-party settings (Gu et al., 2021; Ganesh et al., 2023), where a moderator's decisions regarding interventions and turn assignment (Hydén and Bülow, 2003; Gibson, 2003; Ouchi and Tsuboi, 2016) must account for the conversation context, group dynamics, and the balance of participation. Depending on the scenario, moderators fulfill various functional roles, such as providing background information, facilitating topic transitions, and posing questions to guide discussions and maintain their quality (Wright, 2009; Park et al., 2012). Furthermore, moderators often operate under hybrid motives, which include facilitating quality arguments (Landwehr, 2014), maintaining social engagement (Myers, 2014), and managing external factors like time constraints (Wright, 2009). Ultimately, moderation is a strategic task, requiring the application of specific strategies to encourage constructive contributions and participant engagement while minimizing destructive conflicts (Hsieh and Tsai, 2012; Edwards, 2002; Forester, 2006).

The effect and influence of human conversational moderation have been studied across various domains using different analytical measures. In online mental health support forums, the presence of a moderator has been shown to improve user engagement, openness, linguistic coordination, and trust-building compared to non-moderated groups (Wadden et al., 2021). In the educational domain, moderators have been found to enhance collaboration patterns and increase online participation rates in group learning settings (Hsieh and Tsai, 2012). Case studies and interviews have also been conducted to analyze the role and function of moderators in community building (Cullen and Kairam, 2022; Seering et al., 2019), focus group discussions (Grønkjær et al., 2011), and online public issue discussions and debates (Wright, 2009; Edwards, 2002), mediating contentious stakeholders (Forester, 2006).

Despite the existence of some annotation protocols and datasets, resources for conversational moderation remain notably limited. Many studies have been conducted on small sample sizes (Vasodavan et al., 2020; Hsieh and Tsai, 2012) and often do not make their datasets publicly available (Grønkjær et al., 2011; Wadden et al., 2021). Additionally, some research relies on methodologies such as interviews or case studies, which are not reusable for further analysis or automation (Forester, 2006). The only well-annotated dataset currently available consists of just 300 comments (Park et al., 2012). Furthermore, some studies treat moderation as a reactive intervention to participant comments, and therefore structuring data as comment-intervention pairs (Falk et al., 2024; Grønkjær et al., 2011), thereby overlooking broader session-level objectives such as balancing participation and the overall role of the moderator. Moreover, while several annotation protocols exist, they tend to be overly specific to their application domains. For instance, the role of "resolving site use issues" is only pertinent to e-rule-making scenarios (Park et al., 2012).

3 The WHoW Conversational Moderation Analytic Framework

We designed an analytic framework that (1) is grounded in the existing literature(Park et al., 2012; Vasodavan et al., 2020; Wright, 2009); (2) captures the multifaceted nature of conversational moderation; and (3) generalizes across various dialogue domains. Our final framework (Table 1) is structured around three core dimensions: motives (Why), target speakers (Who), and dialogue acts (How). In addition to dialogue acts, which are widely employed to study dialogue patterns(Shriberg et al., 2004), we incorporated the motive dimension to provide insights into the objectives the moderator seeks to facilitate within a given scenario and context. Furthermore, we introduced the target speaker dimension to explore the moderator's interactive style and strategies for balancing participation in a multi-party setting(Gibson, 2003; Hydén and Bülow, 2003). By decomposing the moderation process into these distinct components and analyzing their interplay, the framework enables the characterization of moderator behavior, particularly in terms of varying level of emphasis on motives, functional roles, and rotation strategies. Table 1 shows the definition of the labels under the three dimensions. To derive our labels and ensure compatibility with existing protocols, we categorized all moderation-related typologies identified in Section 2 into motives and dialogue acts, as detailed in Appendix Table 8.

3.1 Motives: Why does the moderator intervene?

The "Why" component examines the motivations behind a moderator's interventions in conversations, focusing on what the moderator aims to facilitate. Existing protocols typically distinguish socially motivated speech – such as "affective strategy" (Hsieh and Tsai, 2012), and "social functions" (Park et al., 2012) – from argument-driven speech. This aligns with the conversational circumplex framework, which categorizes conversational goals along informational and relational dimensions (Yeomans et al., 2022). Furthermore, in the Intelligent Squared Debate Corpus (Zhang et al., 2016) we have observed instances where speech is motivated by meeting rules, such as adherence to time limits. Consequently, we propose three motives driving moderation behaviors: informational, social, and coordinative motives (Table 1, top). Given that previous studies indicate that a single speech can convey multiple motives(Yeomans et al., 2022), and our pilot studies have also observed this tendency, we treat the annotation of this dimension as a multi-labeling task.

3.2 Dialogue Acts: How does the moderator intervene?

The "How" component focuses on analyzing the dialogue acts or the immediate functions of the moderator's interventions. By examining the sequential patterns of these acts, we gain insights into the strategies employed by moderators. The initial set of dialogue acts was derived from the basic labels of the MRDA corpus (Shriberg et al., 2004), which was developed for annotating multi-party meetings. We adapted the "Question" label into the "information elicitation" category, further subdividing it into "Probing" and "Confronting" acts. Similarly, the "Statement" label was adapted into the "information provision" category, which is further divided into "Instruction," "Interpretation," and "Supplement." These two main categories are instrumental in distinguishing the moderator's role as either a "Contributor" or an "Interviewer" (McLafferty, 2004). Additionally, a "Utility" act label is incorporated to account for other types of speech not covered by the primary categories, such as back-

Dimension	Label	Definition
Motives	Informational (IM)	Provide or acquire relevant information to constructively advance the topic or goal of the conversation.
	Coordinative (CM)	Ensure adherence to rules, plans, and broader contextual constraints, such as time and environment.
	Social (SM)	Enhance the social atmosphere and connections among participants by addressing feelings, emotions, and interpersonal dynamics within the group.
Dialogue acts	Probing (prob)	Prompt speaker for responses.
	Confronting (conf)	Prompt one speaker to response or engage with another speaker's statement, question or opinion.
	Instruction (inst)	Explicitly command, influence, halt, or shape the immediate behavior of the recipients.
	Interpretation (inte)	Clarify, reframe, summarize, paraphrase, or make connection to earlier conversation content.
	Supplement (supp)	Enrich the conversation by supplementing details or information without immediately changing the target speaker's behavior.
	Utility (util)	All other unspecified acts.
Target speaker	Target speaker (TS)	The group or person addressed by the moderator.

Table 1: Definitions and acronyms for the labels across the three dimensions: motives (Why), dialogue acts (How), and target speakers (Who). Target Speaker is a categorical variable with values corresponding to each participant in the dialogue, plus "audience," "self," "everyone," "support side," "against side," "all speakers", and "unknkown".

channeling, floor grabbing, and greetings. The definitions of the labels are included in Table 1. Appendix Table 9 presents example sentences that intersect between the motives and dialogue acts dimensions. We treat dialogue acts as mutually exclusive and formalize it as a sentence level multi-class classification task.

Information elicitation We seek to investigate whether moderators facilitates a conversation by fostering engagement among participants. To this end, we categorize information elicitation behaviors into two types: **Probing** (prompting a speaker to contribute information) and **Confronting** (prompting one speaker to engage with another). This distinction provides insights into whether the moderator gathers information through direct prompts or by fostering interaction between participants.

Information provision Moderators contribute information in order to fill knowledge gaps, manage participants or clarify previous statements (Park et al., 2012; Wright, 2009). We propose three distinct labels for different functions of information provision. **Instruction** refers to contributions intended to immediately alter the behavior of the target speaker. **Interpretation** captures interventions that refer to back to the conversation history, such as summarization. **Supplement** responses provide additional information, including proposals, opinions, and external knowledge.

3.3 Target Speaker: Who does the moderator address?

The "Who" component focuses on identifying the intended target of the moderator's intervention, which differs from the typical task of "next speaker prediction" in multi-party dialogues (Ishii et al., 2019). Since the target participants are not always the subsequent speakers, analyzing the discrepancies between the prior speaker, target speaker, and next speaker enables an assessment of the intended rotations in participation and the moderator's initiatives during the discussion session. We approach the annotation of this dimension as a multi-class classification task, with labels corresponding to the speakers' names and introduce general additional classes, including "everyone", "unknown," and "all speakers", and dataset-specific classes, including "audience", "against team", and "support team".

4 Dataset and Human Annotation

4.1 Data sets

We used the Intelligence Squared Debates (INSQ) Corpus(Zhang et al., 2016), a collection of transcripts from a live-recorded U.S. television debate

	INSQ			RTRP	
	Test	**Dev**	**Train**	**Test**	**Train**
Episodes	19	11	78	20	68
Speakers / episode (Min, Mean, Max)	(4, 4.63, 6)	(4, 4.546, 6)	(4, 4.615, 6)	(2, 3.450, 5)	(3, 4.471, 7)
M share / episode (%)	38%	36%	37%	41%	40%
M Turns/episode	69	73	70	17	21
M Sentences (Total)	2,795	1,702	11,153	1,160	4,341

Table 2: Descriptive statistics for the INSQ and RTRP. M = Moderator. Share is the proportion of words uttered by the moderator. Turn are contiguous (multi-sentence) contribution of the moderator.

	DA	**IM**	**CM**	**SM**	**TS**
INSQ	0.49	0.43	0.37	0.41	0.72
RTRP	0.59	0.67	0.54	0.63	0.75

Table 3: Inter-annotator agreement (Krippendorff's alpha), across the dialogue acts (DA), motives (IM, CM, SM), and target speaker (TS) dimensions for the datasets INSQ and RTRP.

show featuring Oxford-style debates. The corpus comprises 108 episodes covering a wide range of topics, from foreign policy to the benefits of organic foods. Each debate includes a moderator and two teams of experts arguing, respectively, "for" and "against" the topic. Although the debates are structured into three phases (introduction, discussion, and conclusion) our analysis focused exclusively on the interactive discussion phase. The corpus includes information about each speaker's role (moderator, team member, audience member). We randomly split the data into 11 development, 19 test and 78 training episodes.

To validate the generalizability of our framework across scenarios, we expanded our dataset with a subset of The NPR Interview Corpus (Majumder et al., 2020) (RTRP). We specifically selected episodes from a panel discussion program titled "Roundtable," in which the moderator accounts for 30% - 50% of the dialogue, and which involve more than three speakers. This subset features panel discussions with speakers holding diverse views, though not necessarily opposing each other, as in the INSQ data. This selection yielded 88 episodes, from which we randomly sampled 20 to create a test set. Table 2 presents the descriptive statistics of the two subsets.

4.2 Human annotation process

The annotation labeled each sentence of the moderation speech transcript according for the Why, How, and Who dimensions, as illustrated in Figure 1. The development of the annotation schema commenced with two rounds of pilot studies involving authors and NLP PhD students for testing the concept definitions with one episode from each dataset, which resulted in the transition of motive labeling to a multi-label task and a reduction in dialogue act classes from eight to six. For the final annotation phase, we recruited five annotators, all proficient or native English speakers and sudents of linguistics or NLP. We paid at a rate of 36.04 USD/hour, far exceeding the local minimum wage. The annotators manually annotated the development and test sets from the Intelligence Squared Corpus, and the test set of the Roundtable episodes. Annotators received the definitions of labels as outlined in section 3 and Table 1. To facilitate the handling of the multi-classs dialogue act annotation, we developed a decision tree flowchart to achieve consistent prioritisation of labels (see Appendix Figure 4). We conducted one practice annotation round including group discussions to clarify any misconceptions and two further meetings during the annotation phase to discuss remaining misunderstandings. Details of annotation material and interface are provided in Appendix Section F

Each sentence in the moderators' speech transcripts was annotated for the presence of three motives, one identified dialogue act, and the speaker(s) targetted by the intervention. Each episode was annotated by at least two annotators to assess reliability. The finalised annotations were aggregated using majority vote; in cases of evenly divided opinions, the author made the final decision. The resulting inter-annotator agreement, measured by Krippendorff's alpha, is presented in Table 3. The RTRP subset consistently scored higher agreement than the INSQ subset, while overall the agreement ranged from moderate to strong as expected for a complex and nuanced task and consistent with

Model	DA	IM	CM	SM	TS
Random	0.153	0.492	0.508	0.405	0.057
GPT-4o-MT(INSQ)	0.485	0.761	0.711	0.767	0.497
GPT-4o-ST(INSQ)	**0.515**	0.7287	0.686	0.668	**0.525**
longformer-MT(INSQ)	0.494	0.764	0.719	**0.784**	0.246
longformer-ST(INSQ)	0.493	**0.772**	**0.726**	0.694	0.299
GPT-4o-MT(RTRP)	**0.504**	0.726	0.732	**0.754**	**0.467**
GPT-4o-ST(RTRP)	0.492	0.747	0.639	0.635	0.464
longformer-MT(RTRP)	0.414	0.753	**0.774**	0.731	0.196
longformer-ST(RTRP)	0.417	**0.757**	0.759	0.729	0.225

Table 4: Macro-F1 comparing GPT-4o and Longformer using multi-task (MT) and single-task (ST) approaches across the two subsets. The bold numbers highlights the top performer of the dimension in the subset. The random baseline is derived from five random simulations.

Model	DA	IM	CM	SM	TS
MT (INSQ)	0.38	**0.52**	**0.42**	**0.53**	0.66
ST (INSQ)	**0.53**	0.46	0.37	0.34	**0.68**
MT (RTRP)	**0.53**	0.45	**0.51**	**0.46**	0.60
ST (RTRP)	**0.53**	**0.49**	0.28	0.27	**0.61**

Table 5: Krippendorff's alpha agreement between human labels and GPT-4o predictions using single task (ST) or multi-task (MT) prompts for the two datasets.

previous studies s(Falk et al., 2024). A detailed analysis of disagreements is provided in Appendix section C.

5 Automatic Annotation

To develop an automated annotation and analysis pipeline, we utilized the GPT-4o API (OpenAI, 2024) and optimized prompts using the development set from the INSQ subset (details of the prompt structure are provided in Appendix Section B). The pipeline involves five classification tasks: two multi-class classifications for dialogue acts and target speakers, and three binary classifications for motive labels. In addition to predicting each task separately, we experimented with an aggregated multi-task prompt (details in Appendix Figure 5) to predict all tasks simultaneously. The evaluation results of these two approaches, tested on human-annotated datasets, are presented in Table 4. We also measured the agreement between GPT-4o annotations and aggregated human labels using Krippendorff's alpha, as shown in Table 5. Overall, the multi-tasking approach demonstrated greater consistency, with higher average Macro-F1 (0.64 vs. 0.61) and agreement (0.51 vs. 0.46) across dimensions and subsets. As a result, we selected the multi-tasking approach for annotating the training sets. While these scores do not indicate perfect alignment with human annotators, the moderate F1 score demonstrates that the model effectively captures key patterns and distinctions across most dimensions. Additionally, our detailed error analysis in Appendix Section D indicates that most misclassifications arise from subjective interpretations, context dependency, or ambiguity.

5.1 Model comparison

We further validated our automated labels by fine-tuning Longformer models (Beltagy et al., 2020). We compared both single-task and multi-task variants of the Longformer, employing individuals and combined loss functions, respectively (details in Appendix E. The results on the human-labeled test set are presented in Table 4. While the fine-tuned Longformer demonstrated performance comparable to GPT-4o across most dimensions, a notable disparity was observed in predicting the target speaker. This discrepancy may be attributed to the dynamic nature of classification labels across episodes where the number and nature of labels (aka speakers) change between episodes. Generative or retrieval approaches are more effective for target speaker classification.

6 Analysis

We conducted analysis to investigate which dialogue acts are employed to achieve the objectives of discussion facilitation and participation. We investigate specifically how speaker rotation and interaction are facilitated across the two scenarios covered in our dataset. The analysis is based on our full data set, comprising GPT-4o labeled data.

6.1 Motives and Dialogue Acts

Table 6 presents the distributions of the three motives and the six dialogue acts across the two datasets, along with the conditional probabilities of the dialogue acts given the motives, revealing differences between the two settings in terms of functional role, motive prioritization, and strategies employed to achieve these motives.

There is a distinct difference in motive emphasis between the two settings. In INSQ, the moderator is primarily focused on coordinating the discussion process (65%), followed by facilitating the

	prob	conf	inst	inte	supp	util	Total
INSQ							
IM	**0.41**	0.23	0.04	0.12	0.19	0.01	6378
CM	0.15	0.10	**0.54**	0.02	0.09	0.11	10236
SM	0.08	0.02	0.10	0.02	0.14	**0.65**	1724
Total	3606	1864	5497	832	1745	2159	15703
RTRP							
IM	0.41	0.03	0.01	0.03	**0.51**	<0.01	4160
CM	0.06	0.02	**0.43**	0.01	0.33	0.17	1482
SM	0.06	0.01	0.02	0.01	0.28	**0.62**	944
Total	1750	124	633	121	2400	806	5834

Table 6: Comparing the distributions of three motives (rows) and six dialogue acts (cols) across the two data sets. Cells show the conditional probabilities of dialogue acts given the motives. We bold the most common dialogue act per motive, and underline the second most common.

exchange and contribution of information among participants (41%). In contrast, the RTRP moderators are primarily information motivated (71%).

Regarding the differences in functional roles, within the INSQ setting, the moderator predominantly focuses on providing instructions (35%) and probing (22%) speakers to elicit contributions. In contrast, in the RTRP setting, the moderator's primary function is to supply information (41%), complemented by a secondary focus on probing (29%) to invite contributions.

Strategically, **Informational motives** in the INSQ setting are implemented by actively facilitating participant contributions through methods such as probing (0.41) and confronting (0.23), along with notable uses of interpreting (0.12) and supplementing (0.19) information to enhance collective understanding. In contrast, the RTRP setting is characterized by moderators predominantly delivering information themselves (0.51) and engaging participants through probing (0.41). The minimal use of confrontation (0.02) and interpretation (0.01) in RTRP suggests a relatively low emphasis on facilitating interaction among participants.

Coordination motives in both settings primarily rely on instructions. However, INSQ moderators are more likely to engage in coordination through inquiry (0.15), maintaining dialogue engagement by asking participants about their preferences for rotation and participation. In contrast, RTRP moderators tend to provide coordinative information (0.33), such as explaining rules. Social motives in both settings are mainly expressed through "utility" acts, like greetings. Notably, RTRP moderators are more inclined to share social information (0.28 vs. 0.14), such as personal stories, which indicates a more social atmosphere in this setting. While our observations can be partially explained by the respective rules of the discussion programs, they do highlight different high-level strategies to facilitate constructive discussion.

6.2 Balancing Speaker Participation

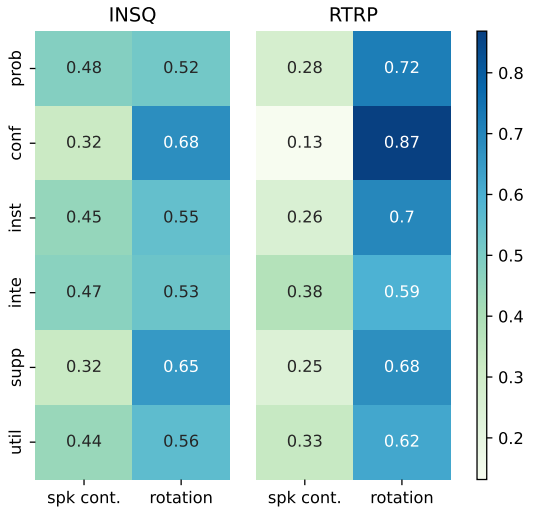


Figure 2: Probabilities of participants' rotation statuses following different moderation dialogue acts.

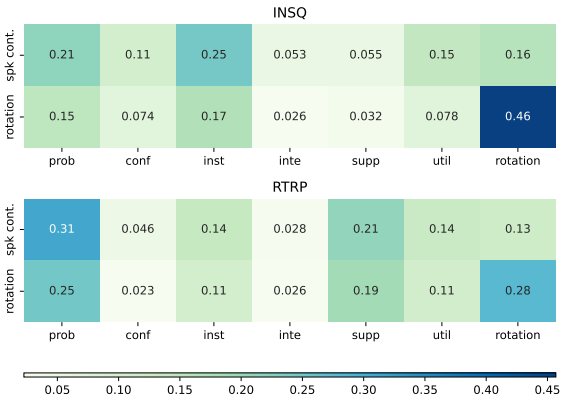


Figure 3: Probabilities of rotation statuses and moderator interventions following different rotation statuses.

An essential role of a moderator is to facilitate balanced participation among participants and their respective stances. To analyze how moderators balance participation, we examine the transition probabilities between moderator dialogue acts (DA) and speaker rotation. We aggregating sentence-level

	Responding	Responded	Specific
INSQ	0.26	0.43	0.49
RTRP	0.20	0.50	0.56

Table 7: Proportion of moderator sentences that respond to, are responded, or are directed at a specific speaker.

DA labels into a turn-level moderator DA. Participants' responses are categorized based on their match with the last non-moderator speaker in the dialogue: a response is labeled as "continuation" if the same speaker continues, and as "rotation" if the speaker changes.[2] We encoded all conversation sequences from the two subsets to construct two matrices to examine the transition between moderators DA and speaker rotation.

Figure 2 provides insights into the use of various dialogue acts for either rotating or continuing speakers in the two data sets. In RTRP, moderators predominantly favor rotating speakers after intervention, whereas INSQ displays a more balanced pattern, albeit with a slightly higher overall tendency to rotate. Among the dialogue acts, confrontation (0.68 & 0.87) consistently has the highest probability of leading to rotation in both datasets. Conversely, interpretation (0.47 & 0.38) is frequently employed to continue the conversation with the same speaker across both scenarios.

Figure 3 shows how and whether moderators would intervene after a speaker finishes. In both settings, likelihood of interventions increase when a speaker continues for multiple exchanges. INSQ moderators primarily use probing and instruction, while RTRP moderators combine these with information supplementation. The 'rotation to rotation' probability, which indicates natural speaker transitions without moderator input, suggests INSQ moderators (0.48 vs. 0.28) are more proactive in facilitating inter-speaker interactions than their RTRP counterparts.

6.3 Pro-activity, Interactivity, and Specificity

By analyzing whether the moderator's target speaker aligns with the speakers before and after their intervention, we can infer the moderator's interaction style in terms of proactivity (initiating vs. responding), interactivity (eliciting a response vs. no response), and specificity (addressing an individual vs. everyone). Table 7 shows that moderators in both settings predominantly engage in proactive interventions rather than passive replies, with moderate levels of interactivity and specificity. Overall, RTRP moderators display higher levels of proactivity, interactivity, and specificity compared to INSQ moderators.

[2]For example, in the conversation depicted in Figure 1, if Russell had conversed with the moderator for more than one exchange, this segment would be encoded as $[continuation, \{prob, inst, supp, util\}, rotation]$.

7 Conclusion

In this study, we developed a automatic analytic framework capable of characterizing conversational moderation across scenarios. This framework dissects the complexity of moderation decision-making into three key decisions: motives, target speaker, and dialogue acts. Using this framework, we annotated moderation speech within two distinct settings: the Intelligent Squared Debate Corpus(INSQ)and the RoundTable Radio Panel Discussion(RTRP). The constructed dataset, which includes 5,657 human-annotated moderation sentences and 15,494 GPT-4o annotated sentences. Additionally, we assessed the feasibility of fine-tuning the Longformer model using the GPT-4o annotated training set.

Our analysis demonstrates the framework's effectiveness in differentiating intervention strategies and styles across various scenarios. In the INSQ setting, moderators are characterized as being more coordination-motivated, playing functional roles as interviewers and instructors, while occasionally facilitating interaction between speakers. In contrast, moderators in the RTRP setting are more information-oriented, taking on both contributor and interviewer roles, as they often contribute to the discussion topics themselves. Although they seek information from the speakers, they rarely facilitate interactions between the participants.

Future studies should encompass a broader range of moderation scenarios datasets, such as group counseling (Kissil, 2016) and second language group conversation (Gao et al., 2024). Moreover, the proposed analytic framework could be expanded to facilitate conversational moderation generation by sequentially predicting the three key components. Finally, there is a need to develop evaluation metrics that assess the effects and biases of moderation interventions (Spada and Vreeland, 2013). This would enable a deeper understanding and optimization of the impact and fairness of moderation practice.

8 Limitations

This study has a few limitations. Some dimensions exhibit low to moderate inter-annotator agreement and low macro-F1 scores, indicating that the boundaries between certain concepts can be ambiguous and subjective. This issue is not unique to our research, as previous studies on moderation-related annotations have also reported both low(Falk et al., 2024) and high(Park et al., 2012) levels of inter-annotator agreement. As shown in Table 3, the agreement levels and macro-F1 scores differ across the settings we analyzed, suggesting that ambiguity is highly context-dependent, with some contexts using more explicit language and others relying on implicit expressions. We recommend that future studies adapting this framework incorporate some degree of human validation tailored to the specific context. Additionally, while we aimed to develop and validate an analytic framework that generalizes across scenarios, the two selected scenarios share a high degree of similarity, both placing less emphasis on social motives. This limitation was due to the lack of sufficient data to compare more diverse scenarios, as multi-party conversation data with clearly tagged moderators are scarce. However, despite the similarity between the selected scenarios, the framework successfully differentiated the two settings, demonstrating its potential for comparative analysis.

9 Ethics Statement

This study was conducted in accordance with the ACL Code of Ethics. Given that the multi-party discussion transcripts may involve controversial topics, annotators were informed in advance and were granted the right to skip any content they found uncomfortable. All identifyable personal information have been removed from the datasets. The annotation protocol and material were approved by local human research ethics committee.

In terms of potential risks and dangers, our work at this stage is primarily analytical and does not involve content generation, thereby minimizing the risk of producing harmful material. Additionally, since the research focuses on moderation rather than persuasion, the findings are unlikely to contribute to harmful uses, such as the spread of propaganda.

10 Instruction for Participants

Acknowledgments

References

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Amanda LL Cullen and Sanjay R Kairam. 2022. Practicing moderation: Community moderation as reflective practice. *Proceedings of the ACM on Human-computer Interaction*, 6(CSCW1):1–32.

Cristian Danescu-Niculescu-Mizil, Lillian Lee, Bo Pang, and Jon Kleinberg. 2012. Echoes of power: Language effects and power differences in social interaction. In *Proceedings of the 21st international conference on World Wide Web*, pages 699–708.

Arthur R Edwards. 2002. The moderator as an emerging democratic intermediary: The role of the moderator in internet discussions about public issues. *Information polity*, 7(1):3–20.

Neele Falk, Eva Maria Vecchi, Iman Jundi, and Gabriella Lapesa. 2024. Moderation in the wild: Investigating user-driven moderation in online discussions. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 992–1013.

John Forester. 2006. Making participation work when interests conflict: Moving from facilitating dialogue and moderating debate to mediating negotiations. *Journal of the American Planning Association*, 72(4):447–456.

Dennis Friess and Christiane Eilders. 2015. A systematic review of online deliberation research. *Policy & Internet*, 7(3):319–339.

Ananya Ganesh, Martha Palmer, and Katharina Kann. 2023. A survey of challenges and methods in the computational modeling of multi-party dialog. In *Proceedings of the 5th Workshop on NLP for Conversational AI (NLP4ConvAI 2023)*, pages 140–154.

Rena Gao, Carsten Roever, and Jey Han Lau. 2024. Interaction matters: An evaluation framework for interactive dialogue assessment on english second language conversations. *arXiv preprint arXiv:2407.06479*.

David R Gibson. 2003. Participation shifts: Order and differentiation in group conversation. *Social forces*, 81(4):1335–1380.

Robert Gorwa, Reuben Binns, and Christian Katzenbach. 2020. Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society*, 7(1):2053951719897945.

James Grimmelmann. 2015. The virtues of moderation. *Yale JL & Tech.*, 17:42.

Mette Grønkjær, Tine Curtis, Charlotte De Crespigny, and Charlotte Delmar. 2011. Analysing group interaction in focus group research: Impact on content and the role of the moderator. *Qualitative studies*, 2(1):16–30.

Jia-Chen Gu, Chongyang Tao, Zhen-Hua Ling, Can Xu, Xiubo Geng, and Daxin Jiang. 2021. Mpc-bert: A pre-trained language model for multi-party conversation understanding. *arXiv preprint arXiv:2106.01541*.

Ya-Hui Hsieh and Chin-Chung Tsai. 2012. The effect of moderator's facilitative strategies on online synchronous discussions. *Computers in Human Behavior*, 28(5):1708–1716.

Lars-Christer Hydén and Pia H Bülow. 2003. Who's talking: drawing conclusions from focus groups—some methodological considerations. *Int. J. Social Research Methodology*, 6(4):305–321.

Ryo Ishii, Kazuhiro Otsuka, Shiro Kumano, Ryuichiro Higashinaka, and Junji Tomita. 2019. Prediction of who will be next speaker and when using mouth-opening pattern in multi-party conversation. *Multimodal Technologies and Interaction*, 3(4):70.

Edward E Jacobs, Robert L Masson, and Riley L Harvill. 1998. *Group counseling: Strategies and skills*. Thomson Brooks/Cole Publishing Co.

Karni Kissil. 2016. About the facilitators. In *The Person of the Therapist Training Model*, pages 77–86. Routledge.

Claudia Landwehr. 2014. Facilitating deliberation: The role of impartial intermediaries in deliberative mini-publics. *Deliberative mini-publics: Involving citizens in the democratic process*, pages 77–92.

Bodhisattwa Prasad Majumder, Shuyang Li, Jianmo Ni, and Julian McAuley. 2020. Interview: Large-scale modeling of media dialog with discourse patterns and knowledge grounding. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8129–8141.

Isabella McLafferty. 2004. Focus group interviews as a data collecting strategy. *Journal of advanced nursing*, 48(2):187–194.

Greg Myers. 2014. Becoming a group: Face and sociability in moderated discussions. In *Discourse and social life*, pages 121–137. Routledge.

OpenAI. 2024. Openai api. OpenAI, https://openai.com/index/hello-gpt-4o/. Accessed: 2024-07-20.

Hiroki Ouchi and Yuta Tsuboi. 2016. Addressee and response selection for multi-party conversation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2133–2143.

Chan Young Park, Julia Mendelsohn, Karthik Radhakrishnan, Kinjal Jain, Tushar Kanakagiri, David Jurgens, and Yulia Tsvetkov. 2021. Detecting community sensitive norm violations in online conversations. *arXiv preprint arXiv:2110.04419*.

Joonsuk Park, Sally Klingel, Claire Cardie, Mary Newhart, Cynthia Farina, and Joan-Josep Vallbé. 2012. Facilitative moderation for online participation in erulemaking. In *Proceedings of the 13th Annual International Conference on Digital Government Research*, pages 173–182.

Joseph Seering, Tony Wang, Jina Yoon, and Geoff Kaufman. 2019. Moderator engagement and community development in the age of algorithms. *New media & society*, 21(7):1417–1443.

Ashish Sharma, Adam S Miner, David C Atkins, and Tim Althoff. 2020. A computational approach to understanding empathy expressed in text-based mental health support. *arXiv preprint arXiv:2009.08441*.

Elizabeth Shriberg, Raj Dhillon, Sonali Bhagat, Jeremy Ang, and Hannah Carvey. 2004. The icsi meeting recorder dialog act (mrda) corpus. In *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue at HLT-NAACL 2004*, pages 97–100.

Paolo Spada and James Raymond Vreeland. 2013. Who moderates the moderators? the effect of non-neutral moderators in deliberative decision making. *Journal of Deliberative Democracy*, 9(2).

Mary Thale. 1989. London debating societies in the 1790s. *The Historical Journal*, 32(1):57–86.

Matthias Trénel. 2009. Facilitation and inclusive deliberation. *Online deliberation: Design, research, and practice*, pages 253–257.

Vinothini Vasodavan, Dorothy DeWitt, Norlidah Alias, and Mariani Md Noh. 2020. E-moderation skills in discussion forums: Patterns of online interactions for knowledge construction. *Pertanika Journal of Social Sciences and Humanities*, 28(4):3025–3045.

Eva Maria Vecchi, Neele Falk, Iman Jundi, and Gabriella Lapesa. 2021. Towards argument mining for social good: A survey. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1338–1352.

David Wadden, Tal August, Qisheng Li, and Tim Althoff. 2021. The effect of moderation on online mental health conversations. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 15, pages 751–763.

Scott Wright. 2009. The role of the moderator: Problems and possibilities for government-run online discussion forums. *Online deliberation: Design, research, and practice*, pages 233–242.

Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2017. Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th international conference on world wide web*, pages 1391–1399.

Michael Yeomans, Maurice E Schweitzer, and Alison Wood Brooks. 2022. The conversational circumplex: Identifying, prioritizing, and pursuing informational and relational motives in conversation. *Current Opinion in Psychology*, 44:293–302.

Justine Zhang, Ravi Kumar, Sujith Ravi, and Cristian Danescu-Niculescu-Mizil. 2016. Conversational flow in oxford-style debates. *arXiv preprint arXiv:1604.03114*.

A Appendix: Framework Supplementary Information

DAs	IM	SM	CM	Source No.
Prob	asking users to provide more infomratino (0), asking user to make or consider possible solution (0), Posing a question at large for the users to respond(0), asking questions (1), asking for elaboration (1), asking for clarification and explanation (1), facilitating students' argumentation (2), conversation stimulator (3)	Empathetic exploration(4)	coordinative enquiry*	0: Park et al. (2012), 1: Vasodavan et al. (2020), 2: Hsieh and Tsai (2012), 3: Wright (2009), 4: Sharma et al. (2020), *: observed from Zhang et al. (2016)
Conf	Encourage users to consider/engage comments of others (0), playing devil's advocate (1), Helping students to sustain threaded discussion (2), Problem Solver (3)	Conflict Resolver (3)	coordinative consensus building*	
Inst	Indicating irrelevant, offpoint comments (0), promote self-regulation (1), Helping students focus on the main topics (2)	invite for team collaboration (1),	directing user to another more relevent issue post more relevent(0), redact and quarantine for inappropriate language content(0), maintaining/encouraging civil deliberative discourse(0), coordinating and planning (1), Open Censor (3), Covert Censor (3), Cleaner (3)	
Inte	Correcting misstatements or clarifying (0), summarisaing discussion (1), highlight contribution (1), archiving information (1), Summarizer of debates (3)	Empathetic interpretation(4)	preference intepretation*	
Supp	Providing information about the proposed rule (0), Pointing to relevant information(0), Pointing out characteristics of effective commenting(0), providing opinion (1), giving feedback (1), introduce other relevant information (1), providing judgment (1), constructive feedback (1), self evaluation (1), Giving students positive feedback (2), Supporter (3), 'Cybrarian' (3)	informal talk (1), adding personal experience/opinion (1), Welcomer (3), Empathetic reaction(4)	explaining the goals/rules of moderation(0), explaining the role of CeRI(0), explaining why comment is outside scope (0),	
Util	acknowledgement*	greeting (1), appreciation (1), humor (1), use emojis (1), making people feel welcome(3),	floor grabbing*	

Table 8: This table presents a collection of literature with taxonomies for moderation, mapping their classifications across the dialogue acts and motives dimensions of our framework.

DAs	IM	CM	SM
Prob	Can you take that on? (prompting) As long as the political spectrum is covered overall, what's wrong with that? (follow up question) Siva? (name calling prompt)	Which of you would like to go first? (preference inquiry) Did this gentleman come down yet? (coordinative question) It's working, right? (question managing environment)	Is that a relief to you or– (asking feeling) Could you tell us your name, please? (social question) Do you have eyeglasses? (humour question)
Conf	That landed pretty well I think, so can you respond to that? (counter confronting) On this side, do you want to respond, or do you agree? (consensus confronting) You actually asked a perfect question, and so Mark Zandi, do you want to take that on? (confronting question)	The other side care to respond, if not I'll move on.(coordinative consensus) Response from the other side, or do you want to pass? (coordinative confronting) Marc Thiessen, do you want to join your partner on this one, because I think– (coordinative consensus)	Bryan Caplan, I think he just described your fantasy, come true.(social confronting) I'd love to hear your answer to that question, so go for it. (confronting with affective appeal) Jared Bernstein, the guy you called "nuts" just said you're unfair. (humour confronting)
Inst	Can you frame your question as a question? (articulate instruction) Relate that point to this motion. (back to topic) I want to stay on the merits of the Obama plan. (manage topic)	Remember, about 30 seconds is what you'll get. (time control) Can you go up three steps, please, and turn right? (coordinating instruction) I'll be right back after this message. (program management)	Do not be afraid. (emotion instruction) Those who agree, just a round of applause to that. (pro-social instruction) –because it's turning into a personal attack. (stop anti-social)
Inte	So, Matt, you're saying that it's not true that it's inevitable that Amazon will control everything. (summarisation) Their point is that it would be a bad thing. (simplification) But that would be the question of mobility. (reframe)	That was an ambiguous signal. (situation interpretation) You're pointing to Lawrence Korb.(preference interpretation) And you want the side arguing for the motion to address that (preference interpretation)	I think it was a rhetorical question, and it got a good laugh. (humour interpretation) And it's a little bit insulting almost to say (toxicity interpretation) —honestly, I don't think that was an—a personal attack— (toxicity interpretation)
Supp	I agree that it is.(agreement) The fact is that one of the US manufacturers, with 1 percent of its yearly production, would run us out of the whole market.(add information) They had never paid any attention whatsoever to Africa. (share opinion)	Fifty-one of you voted against the motion. (vote reporting) And the mic's coming down to you. (describe situation) Round two is where the debaters address each other directly (rule explanation)	You have a colorful sleeve. (social chit-chat) I hate to reward it but I'm going to. (encouragement) And I think all of us probably share a sense that we want things to improve. (state common feeling)
Util	Fair question. (acknoledgement) Right (acknoledgement) So the– (floor grabbing)	All right. (backchanneling) Actually, I– (floor grabbing) Well—(floor grabbing)	Thank you Evgeny Morozov. (thanks) I'm sorry. (apology) Hi. (greeting)

Table 9: This table presents a collection of exemplar sentences at the intersection of the motives and dialogue acts dimensions.

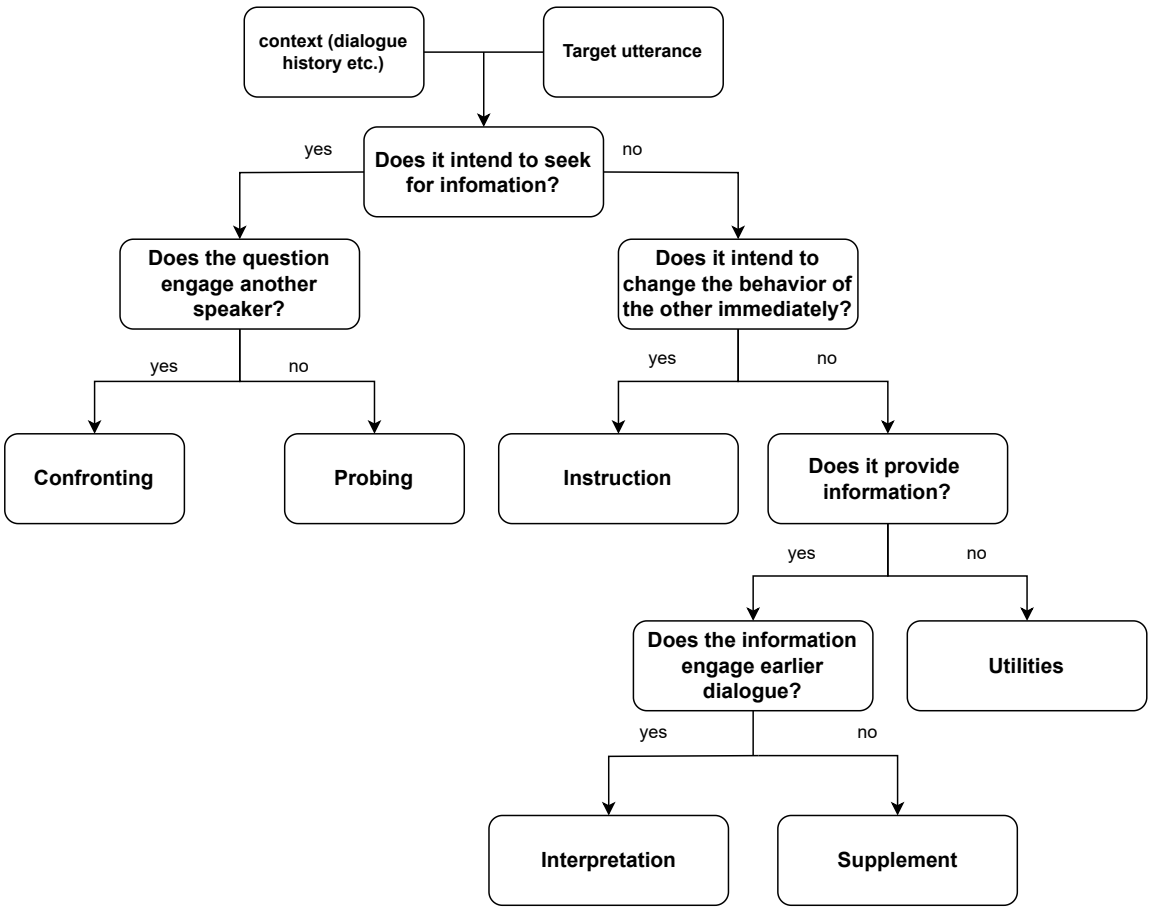


Figure 4: The decision tree used by annotators to resolve ambiguous sentences that may involve multiple dialogue acts.

B Prompt Engineering

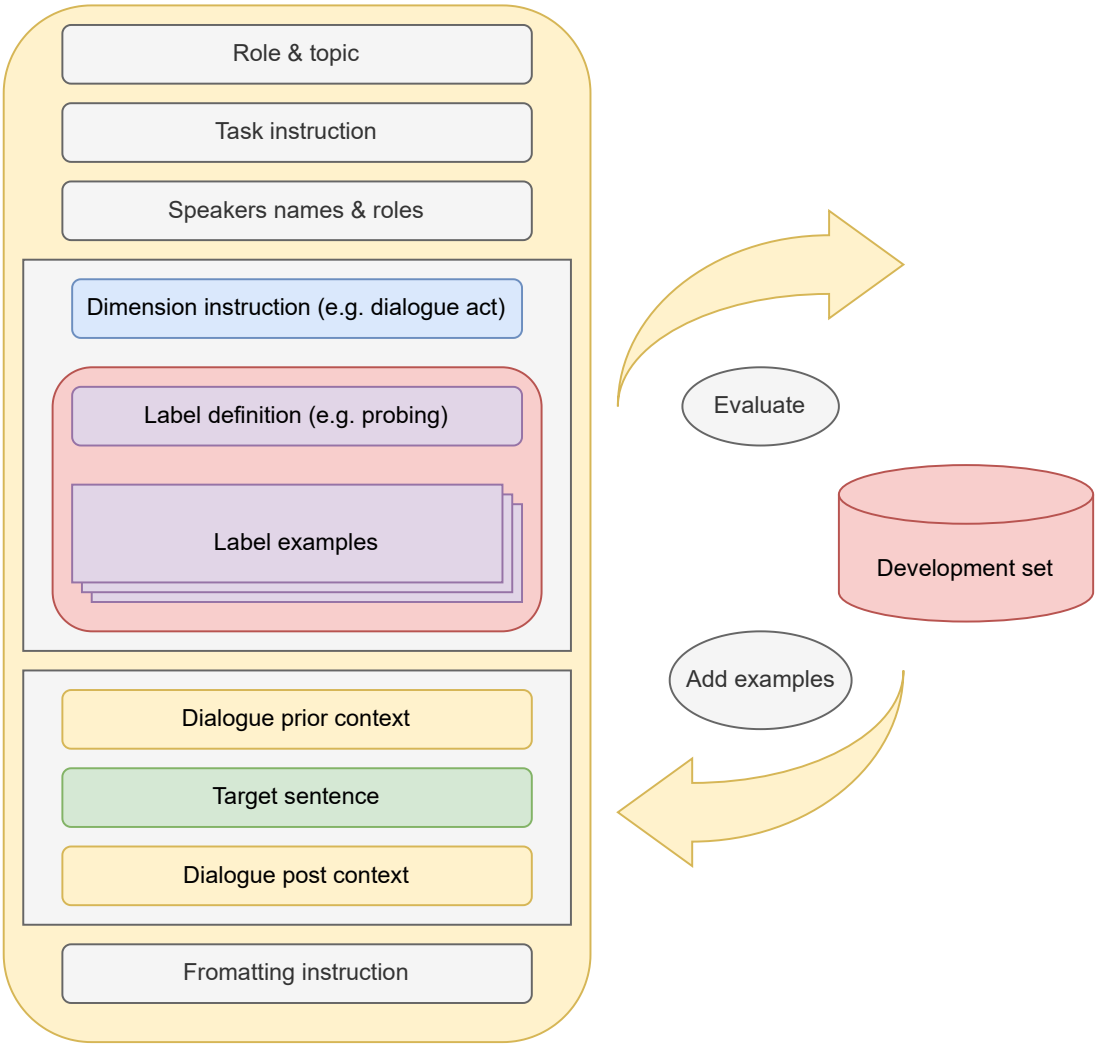


Figure 5: Prompt structure and development cycle

Our prompt design, as illustrated in Figure 5, incorporates several key components: a concise description of the moderation scenario and the annotator's role, an introduction to the task, an explanation of the dimensions and corresponding labels, five preceding responses for context, the target sentence, two subsequent responses for additional context, and instructions for the output format. The label instructions include both definitions from the annotation manual and single-sentence examples. We initially began with a few seed examples for each label and iteratively introduced new examples that had been misclassified during the development process to enhance performance. Table 10 provides a detailed example of a single-task prompt. Additionally, we developed a multi-task prompt that stacks all label definitions and examples across the three dimensions, with adjusted formatting instructions. Table 11 highlights the modifications and stacked elements of the prompt.

section	prompt part
Role & topic	Your role is an annotator, annotating the moderation behavior and speech of a debate TV show. The debate topic is "When It Comes To Politics, The Internet Is Closing Our Minds"
Task instruction	given the definition and the examples, the context of prior and posterior dialogue, please label if the target utterance carries informational motive?
Dimension instruction	Motives: During the dialogue, the moderator is acting upon a mixed-motives scenario, where different motives are expressed through responses depending on the context of the dialogue. Motives are the high level motivation that the moderator aim to achieve. The definitions and examples of the informational motive are below:
Label definition	informational motive: Provide or acquire relevant information to constructively advance the topic or goal of the conversation.
Label examples	examples: “Why do you think minimum wage is unfair?” (Relevant information seeking.) “The legal system has many loopholes.” (Expressing opinion.) “Yea! I agree with your point!” (Agreement relevant to the topic.) “The law was established in 1998.” (Providing topic relevant information.)
Dialogue prior context	Dialogue context before the target sentence: Eli Pariser (for): Just a little story, when I was on the book tour for my book, I was on a radio show in St. Louis. And the host decided to make this big spectacle of having people Google Barack Obama and call-in and read their search results. It was a really boring radio hour. And the first person called in, the second person called in and they interviewed everybody and had people kind of do a read-off where they’re both reading it off at the same time and it was exactly the same. And I was thinking, this is the worse book promotion I’ve ever done. And then a third guy called in, and he said you know it’s the damndest thing, when I Google Barack Obama, the first thing that comes up is this link to this site about how he’s not a natural citizen. And the second link is also a link to a website about how he doesn’t have a birth certificate. Evgeny Morozov (against): That was your publicist. Eli Pariser (for): Oh, I was wondering about that. But so, I think the danger here is that it’s not just that he was getting a view of the world that was really far off the average here. But he didn’t even know that that was the view that he was getting. He had no idea how tilted that view was. And that’s sort of the challenge. I just want to address one other point, which is that there seems to be this question about whether this is happening. And it’s really kind of funny to me, because if you talk to these companies and if you listen to what they’re saying, all of these companies are very clear that personalization is a big part of what they’re doing and what they’re– Evgeny Morozov (against): For pizza, weighted decisions. They are very clear. And they say we don’t want to do it for politics, we only want to do it for pizza. Eli Pariser (for): Right, and the question is, can you trust them? John Donvan (mod): Let me– Jacob, I think Eli left a pretty good image hanging out there, of these folks truly not knowing how much they don’t know and believing what they’re getting and not understanding how slanted it is.
Target sentence	Target sentence: John Donvan (mod): That landed pretty well I think, so can you respond to that?
Dialogue post context	Dialogue context after the target sentence: Jacob Weisberg (against): But a guy who called into a radio show? I know the plural of anecdote is data. But I mean, if this were really happening in the way you say it is, wouldn’t there be some kind of decent study that actually showed widely varying results? I mean as I say, I’ve tried to test this out as best I can. I’ve tried it myself on various browsers, signed in, signed out, Wikipedia always comes up first, sometimes it comes up second. Wikipedia’s vaccine entry is pretty good. I do not think there is actually the kind of variety you’re talking in searches done most of the time by most people. John Donvan (mod): Siva.
Formatting instruction	Please answer only for the target sentence with the JSON format:{"verdict": 0 or 1,"reason": String} For example: answer: {"verdict": 1, "reason": "The moderator asks a question to Joe Smith aimed at eliciting his viewpoint or reaction to a statement from the recent policy change for combatting climate change....."}

Table 10: An example of a single task prompt to determine if the target sentence has informational motive.

section	prompt part
Task instruction	given the definition and the examples, the context of prior and posterior dialogue, please label which motives the target response carries? And which dialogue act the target sentence belong to? And who is the moderator talking to?
Motives section	Motives: During the dialogue, the moderator is acting upon a mixed-motives scenario, where different motives are expressed through responses depending on the context of the dialogue. Different from dialogue act, motives are the high level motivation that the moderator aim to achieve. The definitions and examples of the 3 motives are below: informational motive: Provide or acquire relevant information to constructively advance the topic or goal of the conversation. examples: “Why do you think minimum wage is unfair?” (Relevant information seeking.) “The legal system has many loopholes.” (Expressing opinion.) “Yea! I agree with your point!” (Agreement relevant to the topic.) “The law was established in 1998.” (Providing topic relevant information.) social motive: Enhance the social atmosphere and connections among participants by addressing feelings, emotions, and interpersonal dynamics within the group. examples: “It is sad to hear the news of the tragedy.” (Expressing emotion and feeling.) “Thank you! Mr. Wang.” (Appreciating.) “Hello! Let’s welcome Dr. Frankton.” (Greeting.) “I can understand your struggle being a single mum.” (Empathy) “How do you feel? when your work was totally denied.” (Exploring other’s feeling.) “Please feel free to say your mind because I can’t bite you online, hehe!” (Humour.) “The definition is short and simple! I love it!” (Encouragement.) “Maybe Amy’s intention is different to what you thought, you guys actually believe the same thing.” (Social Reframing.) coordinative motive: Ensure adherence to rules, plans, and broader contextual constraints, such as time and environment. examples: “Let’s move on to the next question due to time running out.” (Command) “We going to start with the blue team and then the red team” (Planning) “Do you want to go first?” (Asking for process preference.) “Please move to the left side and turn on your mic!” (Managing environment)
Dialogue act section	Dialogue act: Dialogue acts is referring to the function of a piece of a speech. The definitions and examples of the 6 motives are below: Probing: Prompt speaker for responses. examples: “What is your view on that Dr. Foster?” (Questioning.) “Where are you from?” (Social questioning.) “Peter!” (Name calling for response.) “If the majority of people are voting against it, would you still insist?” (Elaborated questioning.) “Do you agree with this statement?” (Binary question.) Confronting: Prompt one speaker to response or engage with another speaker’s statement, question or opinion. examples: “So David pointed out the critical weakness of the system, what is your thought on his critiques, Dr. Foster?”, "Judge Anderson, what is your response to this hypothetical scenario posed by Ms. Lee regarding privacy laws?", "Senator Harris, you have proposed reducing taxes instead. How do you respond to Mr. Walkers suggestion to increase school funding?", "So, Dr. Green, Professor Brown just criticized the emissions policy. What is your response to his critique?" Supplement: Enrich the conversation by supplementing details or information without immediately changing the target speaker’s behavior. examples: “And that concludes round one of this Intelligence Squared U.S. debate where our motion is Break up the Big Banks.” (Addressing progess) “The blue team will go first, then the red team can speak” (explaining program rule) “Supposed we live in a world where such behaviour is accepted.” (Hypothesis) “I suggest the best solution is giving everyone equal chances.” (Proposal) “The government announced tax raise from March.” (Providing external information) “I agree with that you said.” (Agreement) “GM means genetic modified.” (Providing external knowledge) “I think people should be given the right to say no!” (Opinion) "The guy with the blue shirt." (Describing appearance) "The power is off." (Describing situation). “In this section, debaters will address one another and also take questions from the audience.” (Explaining upcoming segment) "Let me move this along a little bit further to a slightly different topic, although we have circled around it." (Explaining self intention) "I want to remind you that we are in the question and answer section." (Remind current phase of the discussion) Interpretation: Clarify, reframe, summarize, paraphrase, or make connection to earlier conversation content. examples: “So basically, what Amy said is that they didn’t use the budget efficiently”. (Summarisation) “You said ‘I believe GM is harmless,’.” (Quote) “In another word, you don’t like their plan.”. (Paraphrase) “My understanding is you don’t support this due to moral reason.” (Interpretation) “She does not mean to hurt you but just tell the truth.” (Clarify) “So far, we have Dr. Johnson suggesting. . . ., and Dr. Brown against it because. . . . . ”(Summarisation) “Amy saying that to justify the reduction of the wage, but not aiming to induce suffering.” (Reframing) Instruction: Explicitly command, influence, halt, or shape the immediate behavior of the recipients. examples: “Please get back to the topic.” (Commanding) “Please stop here, we are running out of time.” (Reminding of the rule) “The red will start now.” (Instruction) “Please mind your choice of words and manner.” (social policing) “Do not intentionally create misconception.” (argumentative policing) “Now is not your term, stop here.” (coordinative policing) “What you need to do is raise your hand, and ushers will come to you.” (Guiding participation) “Turn on your microphone before speaking.” (Technical instruction) All Utility: All other unspecified acts. examples: “Thanks, you.” (Greeting) “Sorry.” (Apology) “Okay.” (Back channelling) “Um hm.” (Back channelling) “But, but, but. . . . . ” (Floor grabbing)
Formatting instruction	Please answer only for the target sentence with the JSON format:{"motives": List(None or more from "informational motive", "social motive", "coordinative motive"),"dialogue act": String(one option from "Probing", "Confronting", "Supplement", "Interpretation", "Instruction", "All Utility"),"target speaker(s)": String(one option from "0 (Unknown)", "1 (Self)", "2 (Everyone)", "3 (Audience)", "4 (Eli Pariser- for)", "5 (Siva Vaidhyanathan- for)", "6 (Evgeny Morozov- against)", "7 (Jacob Weisberg- against)", "8 (Support team)", "9 (Against team)", "10 (All speakers)"),"reason": String} For example: answer: {"motive": ["informational motive"], "dialogue act": "Probing", "target speaker(s)": "7 (Joe Smith- for)", "reason": "The moderator asks a question to Joe Smith aimed at eliciting his viewpoint or reaction to a statement from the recent policy change for combatting climate change....."}

Table 11: An example of a multi-task prompt. Here we only demonstrate the components that are different from the single-task prompt.

C Disagreement Cases Analysis

Dimensions	Examples
Dialogue act	1. You know, what do you think about that, Callie? (prob & conf) 2. Our time has run out. (supp & inst) 3. Well let me move on to our final topic, which is gentrification. (supp & inst) 4. Rick MacArthur cited Mexico, it has worked for Mexico.(supp & inte) 5. Yeah. (supp & util)
Motives	6. Can you take that on? (IM vs. CM) 7. Okay, go ahead. (IM vs. CM) 8. Let's let Jacob Weisberg (IM vs. CM) 9. So Lenny took the initiative of sending a question into us by email. (IM vs. SM) 10. Do you agree that our nation needs affirmative action for intelligent conversation? (IM vs. SM) 11. All right. (CM vs. SM)
Target Speaker	12. And that concludes round one of this Intelligence Squared US debate (everyone vs. audience) 13. Let's bring Evgeny in and– (everyone vs. Evgeny) 14. And we also– is Lenny Gengrinovich here? (everyone vs. Lenny)

Table 12: Examples of disagreement cases across the dimensions of dialogue acts, motives, and target speaker. Bracketed information includes the combinations of disagreed labels. All examples are from the INSQ dataset.

In this appendix, we highlight the complexity and difficulty of the task by curating several examples in Table 12. We analyze and discuss cases of disagreement, particularly within the INSQ subset, which received a relatively low agreement score.

To better understand the disagreements in dialogue act annotations, we calculated the co-occurrences of human annotators' votes, as shown in Figure 6. While most dialogue act labels exhibit strong internal consistency, indicating general agreement among annotators, the figure reveals two primary sources of disagreement. The first source involves cases of 'confrontation,' where disagreement often arises when the moderator does not explicitly mention the intended participant by name, leading to differing interpretations of whether the confrontation is implied or direct (Example 1). The second source of disagreement involves the label 'supplement,' which frequently co-occurs with 'instruction,' 'interpretation,' and 'utility.' Examples 2 and 3 illustrate instances where it is unclear whether the moderator is expecting a behavioral change from the recipient or merely providing a reminder or explanation. Additionally, there are numerous ambiguous cases between 'supplement' and 'utility,' such as brief responses like 'Yeah,' where it is uncertain whether the expression is intended as acknowledgment or simple backchanneling.

For disagreements regarding motive labels, we found that the 'coordinative' motive was particularly often confused with the other two categories. Examples 6 to 8 highlight cases where vague probing led some annotators to interpret the moderator's actions as rotating turns according to program rules, while others perceived the probing as an attempt to prompt information from the speakers to contribute to the topic. Short utility phrases like 'All right,' as seen in Example 10, also present ambiguity in motive—whether it's meant for pacing or calming the speaker's emotions is unclear. Additionally, disagreements were noted in the target speaker dimension. In Example 12, it is uncertain whether the moderator is addressing everyone or just the audience. Similarly, in Examples 13 and 14, the addressee shifts mid-sentence, leading to further confusion.

These analyses underscore the inherent complexity and subjectivity involved in labeling dialogue acts and motives. Despite efforts to create clear definitions and guidelines, the nuanced nature of communication often results in differing interpretations among annotators, especially when dealing with implicit intentions, vague statements, or multi-functional phrases.

Figure 6: The normalized co-occurrence matrix of dialogue act human votes from the INSQ subset.

D Classification Error Analysis

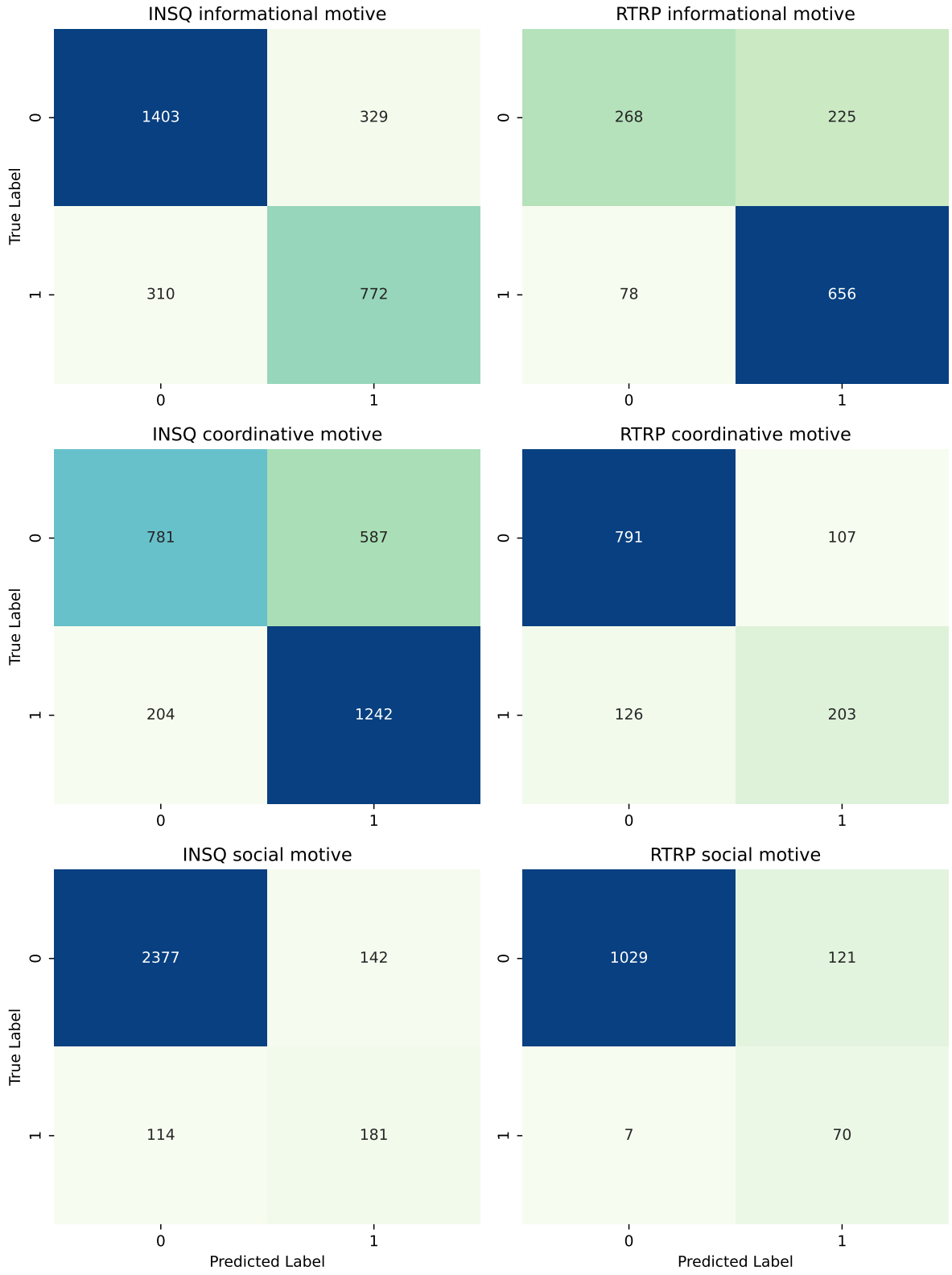


Figure 7: The confusion matrices for the three motives across the two subsets.

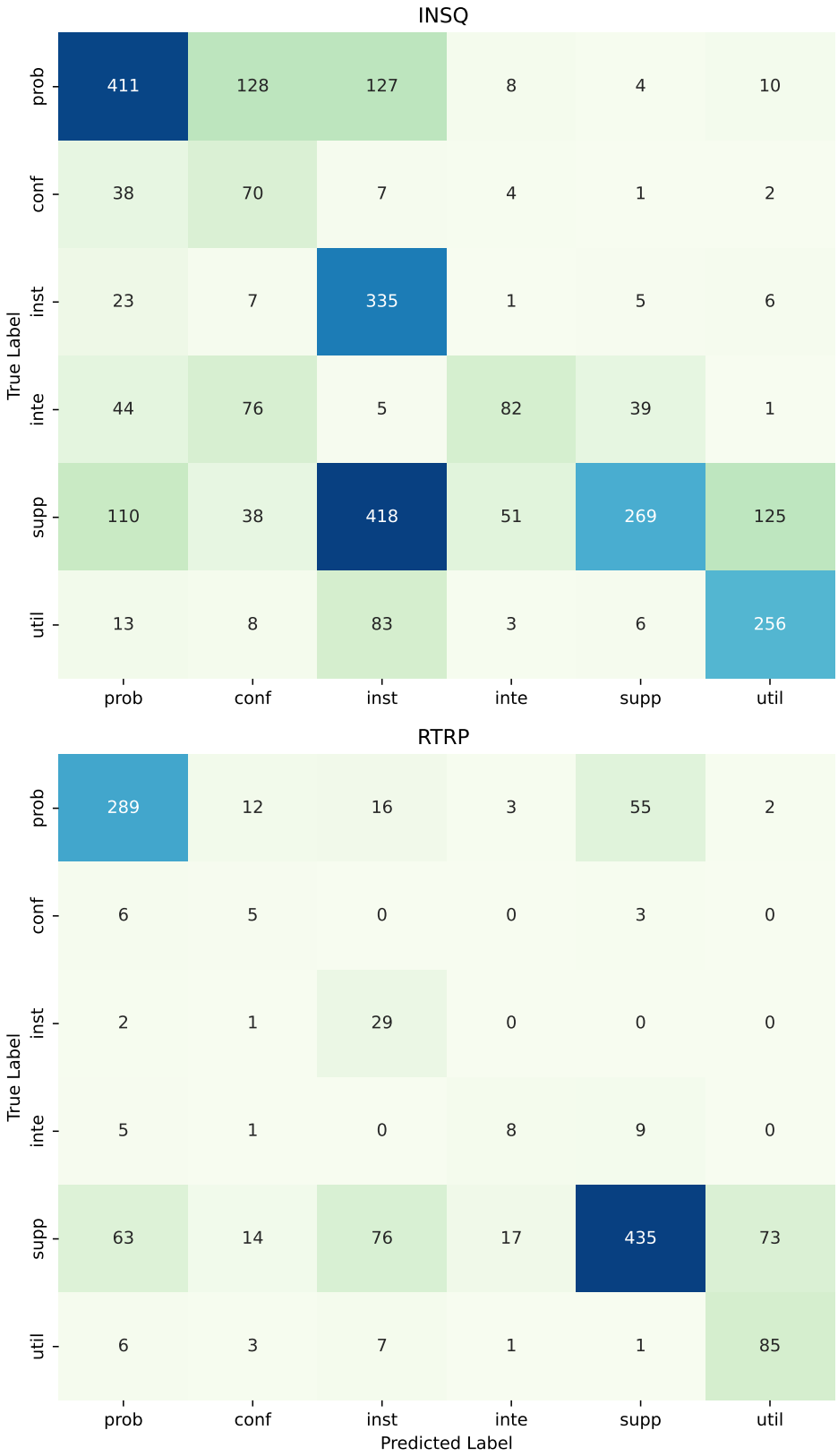


Figure 8: The confusion matrices for the three motives across the two subsets.

Dimensions	Examples
Dialogue act	1. Eli Pariser. (prob vs. conf, INSQ). 2. Dr. David Satcher. (conf vs. prob, INSQ). 3. I want to bring Matt back into this conversation. (prob vs. conf, INSQ) 4. But wasn't your partner using the "that's what happened to me when I typed in Egypt"? (prob vs. inte, INSQ) 5. Let's go to Frank Foer. (prob vs. inst, INSQ) 6. There was a lot of questions that came up during Jena Six, saying, oh, marching is so 1965.(prob vs. supp, RTRP) 7. Your opponents are saying that Amazon cannot be trusted, that it's becoming more and more powerful, and that's probably likely to continue, although you're saying there are mitigating forces.(inte vs. conf, INSQ) 8. Also, that in a peace process that is going nowhere, that is stuck, it lays down a marker that the Israelis cannot ignore.' (inte vs. supp, INSQ) 9. I have a– question in the second row. (supp vs. prob, INSQ) 10. You work for the Washington Post and I couldn't even find the story online about that. (supp vs. prob, RTRP) 11. We're going to ask you to vote again at the end and the team that has moved its numbers the most will be declared our winner. (supp vs. inst, INSQ) 12. Microphones will be brought forward if you raise your hand. (supp vs. inst, RTRO) 13. Yep (supp vs. util, INSQ) 14. Alright (util vs. inst, INSQ)
Motives	15. So how would you relate that directly to the motion? (IM false positive, INSQ) 16. Jacob Weisberg. (IM false negative, INSQ) 17. What do you - Jasmyne, I'll start with you - unfold your, uncross your arms. (IM false negative, RTRP) 18. The team arguing against the motion, Franklin Foer and Scott Turow, they're saying, "It's all a trap. (CM false positive, INSQ) 19.Our motion is "America is to Blame for Mexico's Drug War," at the start, 43 percent of you were for. . . 22 percent against, and 35 percent undecided. (CM false negative, INSQ) 20. Today on our Bloggers' Roundtable, we're taking a close look at urban education and the race for the White House. (CM false positive, RTRP) 21. Well, you're laughing because you think it's impossible or what is. . . (SM false positive, RTRP) 22. All right. (SM false negative, INSQ)
Target Speaker	23. Round two is where the debaters address each other directly and also answer questions from you in the audience and from me. (audience vs. everyone, INSQ) 24. Let me ask the side that's arguing that when it comes to politics, the internet is closing our minds. (support team vs. all speakers, INSQ) 25. But Evgeny kind of addressed that point when he– I think you said, Evgeny, earlier in your opening statements, that initially the theory was the internet gave us tools to do stuff that we were already doing. (audience vs. Evgeny, INSQ) 26. Let me approach this from a couple of different angles. (all speakers vs. audience, RTRP)

Table 13: Examples of error cases across the dimensions of dialogue acts, motives, and target speaker. Bracketed information indicates the predicted labels vs. the human-aggregated labels, along with the source of each example.

In this appendix, we examine the discrepancies between the GPT-4o-based classification results and the aggregated human annotation labels. Figure 7 presents the confusion matrix for the three motives, comparing GPT-4o with the aggregated human annotations, while Figure 8 displays the confusion matrix for the six dialogue act labels. Table 13 provides examples of common errors across the three dimensions to support further qualitative analysis.

An analysis of the dialogue act confusion matrix in Figure 8, particularly within the INSQ subset, reveals four primary sources of error. First, several probing sentences are frequently misclassified as confrontational or instructional. In Table 13, Examples 1 and 2 illustrate instances where the sentences merely include the addressees' names, and the intended purpose of the moderator—to engage the addressees with a previous speaker—depends heavily on the conversational context and remains inherently subjective. Ambiguous cases, such as Example 5, demonstrate scenarios where it is unclear whether the moderator is seeking information or simply inviting someone to participate. Additionally, long

sentences may be reasonably associated with more than one dialogue act, as seen in Example 7, where both interpretation and confrontation are plausible classifications. A substantial number of errors also arise from confusion between 'supplement' and 'instruction,' which is the largest source of misclassifications. In Examples 11 and 12, it is often uncertain whether the moderator is merely explaining or reminding participants of a rule or the program's progress, or if they expect a specific response. Lastly, numerous errors involve brief utility phrases like 'Yep' and 'Alright,' as in Examples 13 and 14. These phrases are highly context-dependent, making it challenging to determine whether the moderator is expressing acknowledgment, signaling the speaker to stop, or simply backchanneling.

Analyzing the confusion matrix for motive prediction in Figure 7, we identified two primary sources of error. In the INSQ subset, the 'coordinative' motive exhibited the lowest performance, with most errors being false positives. For example, in Table 13, Example 18 involves the moderator introducing a key argument for the opposing team. Although this instance was annotated as driven by an informational motive, GPT-4o incorrectly interpretate it as an coordinative move for setting up the introduction. A similar pattern is observed in Example 20 from the RTRP subset, where the moderator introduces the discussion's background and topic. While GPT-4o classified this action as coordination-driven, human annotators labeled it as informational, despite one annotator also indicating a coordinative motive. Additionally, errors related to social motives proved particularly difficult to interpret, as seen in Examples 21 and 22.

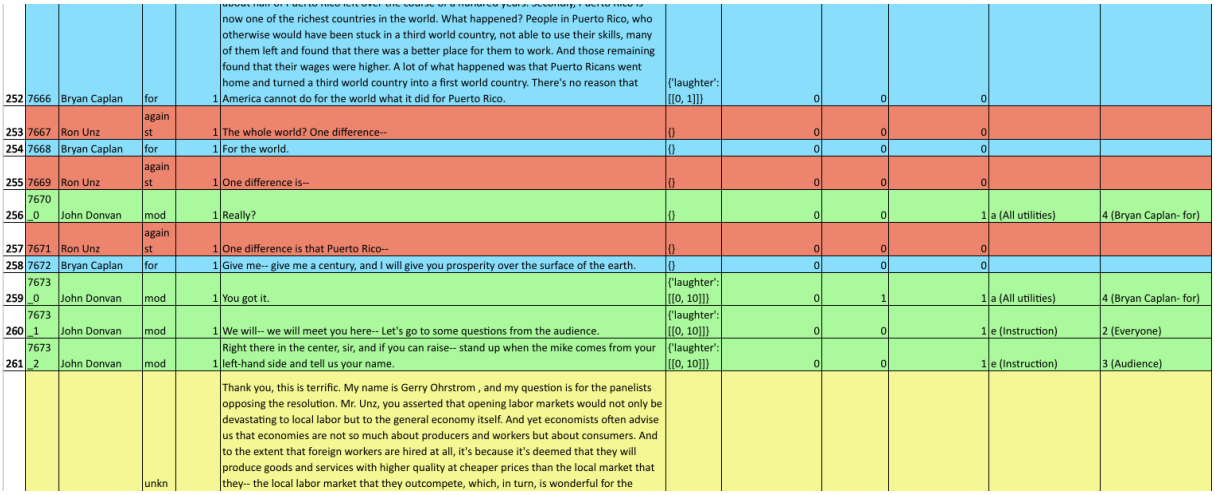
In terms of target speaker classification errors, most misclassification occur when the target speaker is plural,e.g. "everyone". When multiple speakers are addressed, determining the scope or boundary of the intended recipients can be subjective and ambiguous. Examples 23, 24, and 26 illustrate the difficulty in discerning whether the moderator is addressing the entire group or only the audience. Another common source of error arises when the speaker shifts the intended recipient mid-sentence, as demonstrated in Example 25.

In our error case analysis, we identified several instances where GPT-4o classifications diverged from human annotations. However, these misalignments are not always unreasonable. Many examples are highly context-dependent, subjective, and open to interpretation, particularly in cases involving long sentences that could be associated with multiple labels or extremely short sentences, such as name-calling or backchanneling, where interpretation relies heavily on the conversational context. We also examined the reasons generated by GPT-4o to justify its classifications and found that, while they differ from the aggregated human annotations, the majority of these justifications are still defensible.

E Longformer finetuning

We fine-tuned the Hugging Face pre-trained model allenai/longformer-base-4096. The input sequence included the discussion topic, a list of speaker options consisting of all speaker names along with "unknown," "everyone," "audience," and "all speakers." For the INSQ subset, additional speaker options "against team" and "support team" were included. The input also contained the five utterances preceding the target sentence and two utterances following it, with a maximum input length set to 3072 tokens. The model was trained for 3 epochs for 3 hours with a learning rate of 2e-5 using the AdamW optimizer (weight decay = 0.01) and a batch size of 8 on an A100 GPU via the Spartan cluster. For the multi-task approach, we adapted the model to include multiple classifier heads, each corresponding to a different classification task, and backpropagated using a combined loss function.

F Annotator instruction and material



					[illegible] now one of the richest countries in the world. What happened? People in Puerto Rico, who otherwise would have been stuck in a third world country, not able to use their skills, many of them left and found that there was a better place for them to work. And those remaining found that their wages were higher. A lot of what happened was that Puerto Ricans went home and turned a third world country into a first world country. There's no reason that							
252	7666	Bryan Caplan	for	1	America cannot do for the world what it did for Puerto Rico.	{'laughter': [[0, 1]]}	0	0	0			
253	7667	Ron Unz	against	1	The whole world? One difference--	{}	0	0	0			
254	7668	Bryan Caplan	for	1	For the world.	{}	0	0	0			
255	7669	Ron Unz	against	1	One difference is--	{}	0	0	0			
256	7670_0	John Donvan	mod	1	Really?	{}	0	0	1	a (All utilities)	4 (Bryan Caplan- for)	
257	7671	Ron Unz	against	1	One difference is that Puerto Rico--	{}	0	0	0			
258	7672	Bryan Caplan	for	1	Give me-- give me a century, and I will give you prosperity over the surface of the earth.	{}	0	0	0			
259	7673_0	John Donvan	mod	1	You got it.	{'laughter': [[0, 10]]}	0	1	1	a (All utilities)	4 (Bryan Caplan- for)	
260	7673_1	John Donvan	mod	1	We will-- we will meet you here-- Let's go to some questions from the audience.	{'laughter': [[0, 10]]}	0	0	1	e (Instruction)	2 (Everyone)	
261	7673_2	John Donvan	mod	1	Right there in the center, sir, and if you can raise-- stand up when the mike comes from your left-hand side and tell us your name.	{'laughter': [[0, 10]]}	0	0	1	e (Instruction)	3 (Audience)	
			unkn		Thank you, this is terrific. My name is Gerry Ohrstrom , and my question is for the panelists opposing the resolution. Mr. Unz, you asserted that opening labor markets would not only be devastating to local labor but to the general economy itself. And yet economists often advise us that economies are not so much about producers and workers but about consumers. And to the extent that foreign workers are hired at all, it's because it's deemed that they will produce goods and services with higher quality at cheaper prices than the local market that they-- the local labor market that they outcompete, which, in turn, is wonderful for the							

Figure 9: The Excel sheet annotation interface used for annotating moderator transcript.

Exploring the role and behaviour of debate and panel session moderator

PROJECT OVERVIEW

Deliberation is a process of careful and thoughtful discussion, typically involving multiple individuals or stakeholders, to weigh various ideas, viewpoints, arguments, and evidence before making a decision or reaching a conclusion. In real life, deliberative conversation take place in forms of debate, online discussion, parliament meeting etc. While several studies have looked at how to win a debate or argument, extremely few have investigated the role and the functioning of the moderator in facilitating a better conversation between individuals with different point of views. The goal of this research project is improving human deliberative conversation by exploring, analysing, and modelling the behaviours and bias of moderator from existing debate transcripts. We specifically investigate **1) HOW does the moderator did: unveiling patterns in the moderator's interventions and speech, 2) WHY the moderator did these: identifying the motives underpinning these interventions within the context of speaker dialogues,** and **3) WHO are the moderator talking to: investigate the choice of turn assignment and target speaker from the moderator.**

What are the possible benefits?

The project's primary benefit lies in advancing our understanding of moderator behaviours and bias, which serves as a foundation for the development of automatic discussion moderating agents and the detection of moderating bias, which can be used to improve the productivity, efficiency and harmony of human dialogue.

What are the possible risks?

There are no immediate risks that we can foresee, however, due to the nature of debate there might be some controversial, sensitive, and emotional topics and content be exposed to you. but you are free to withdraw from the experiment at any time should you wish to do so. Before the annotation of each debate, we will show you a debriefing including the title, speakers, and the short relevant background information. You may choose to replace the current topic if you feel uncomfortable.

What will happen to information about me?

Regarding data privacy for Mechanical Turk contributors, only internal worker IDs will be accessible to our research team, thereby ensuring that no personally identifiable information is collected. For local participants, essential contact details and payment information will be required; this data will be securely stored on the University of XXX's OneDrive, protected by password encryption until the project's conclusion. Task-related annotated data will also be initially stored on the University of XXX OneDrive.

Prior to any public release, the data will undergo a sanitization process to remove any potential personally identifiable information, ensuring participant privacy is maintained when the data is published in the public repository on GitHub.If you would like more information about the project, please contact the researchers given above.

DATA

Currently, we are expanding the existing **"Intelligence Squared Debates Corpus",** a dataset consisting of full transcripts of debates from the famous American debate TV show with clear labels of speakers' roles and stances (for vs. against). Specifically, we are focusing on the cross-examination phase of the debates, where frequent interactions occur between the moderator and speakers from both sides. In addition, we are also including the transcript from "**Roundtable**" a panel discussion radio show.

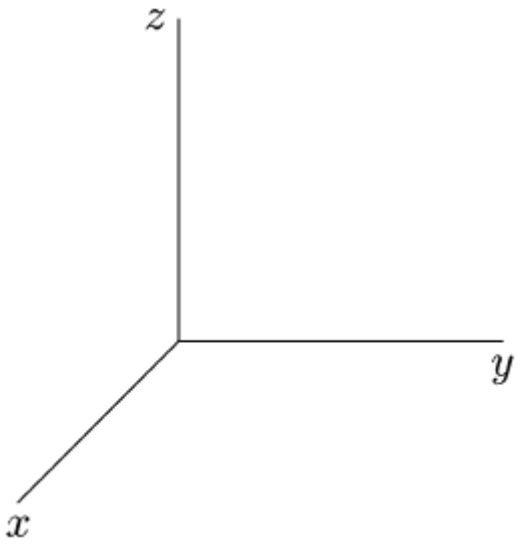
ANNOTATION FACETS AND LABELS INTRODUCTION

For each annotation task, you will be provided some background information, including the topic of the debate, speaker's name and stances, and a segment of complete debate transcript including the interventions from the moderator.

Since we are only interested in moderator's behaviour, you will only need to label the moderator's responses. There are three facets that we would like you to label, which are **motives**, **dialogue acts**, and **target speaker**. At the end of the annotation of each episode, there is also a short survey for your overall impressions of the moderator and the dialogue before and after the annotation.

WHY Motives

In our proposed framework, we assume that the moderator is acting upon a mixed-motives scenario, where different motives are expressed through responses depending on the context of the dialogue. In the framework we proposed, we assume during the debate the moderator wants to achieve informational goals (e.g. argument and knowledge), social goals (e.g. relation building, and stabilising emotion), and coordinating goals (e.g. following rules.):



1.) **Informational Motive (z): Provide or acquire relevant information to constructively advance the topic or goal of the conversation.**.
2.) **Social Motive (x): Enhance the social atmosphere and connections among participants by addressing feelings, emotions, and interpersonal dynamics within the group**.
3.) **Coordinative Motive (y): Ensure adherence to rules, plans, and broader contextual constraints, such as time and environment**.

Based on these assumptions, we identified and proposed three motives dimensions. The definition of each motive dimensions with examples are shown below:

Informational motive (I):

Definition: Provide or acquire relevant information to constructively advance the topic or goal of the conversation..

Examples:

"Why do you think minimum wage is unfair?" (Relevant information seeking.)

"The legal system has many loopholes." (Expressing opinion.)

"Yea! I agree with your point!" (Agreement relevant to the topic.)

"The law was established in 1998." (Providing information.)

Social motive (S):

Definition: **Enhance the social atmosphere and connections among participants by addressing feelings, emotions, and interpersonal dynamics within the group**.

Examples:

"It is sad to hear the news of the tragedy." (Expressing emotion and feeling.)

"Thank you! Mr. Wang." (Appreciating.)

"Hello! Let's welcome Dr. Frankton." (Greeting.)

"I can understand your struggle being a single mum." (Empathy)

"How do you feel? when your work was totally denied." (Exploring other's feeling.)

"Please feel free to say your mind because I can't bite you online, hehe!" (Humour.)

"The definition is short and simple! I love it!" (Encouragement.)

"Maybe Amy's intention is different to what you thought, you guys actually believe the same thing." (Social Reframing.)

Coordinative motive (C):

Definition: **Ensure adherence to rules, plans, and broader contextual constraints, such as time and environment**.

Examples:

"Let's move on to the next question due to time running out." (Command)

"We going to start with the blue team and then the red team" (Planning)

"Do you want to go first?" (Asking for process preference.)

"Please move to the left side and turn on your mic!" (Managing environment)

Mixed motive (I/S/C):

There are also possibilities that one single sentence carries more than one motives.

Example:

"I am very sorry about the incident, but few exceptions cannot defy the statistic majority" (I & S).

"My daughter dies because of a broken traffic light." (I & S).

"Sorry, John, I spoke over you, go ahead?" (S & C)

"Okay—thank you, we—those are good, those are all questions and they're quite good and brief." (I, S & C).

WHAT: Dialogue acts

Dialogue acts is referring to the intention of a piece of dialog. Labelling dialogue act allow us to identify the behaviour pattern and even strategy of the moderator. Based on our observation of the moderator acts, we identified and proposed 3 broad categories and 5 specific acts for as shown below:

Information seeking behaviour:

The goal of the moderator is to facilitate contribution of views, feeling, opinion and knowledge from the participants, therefore information seeking behaviours play a major role in moderation. In addition, we are interested in how moderator foster interaction between the participants, therefore, we separate the information seeking behaviour into two broad categories (probing, confronting) **diverged by if another speaker is linked, engaged or mentioned in the prompt.**

Probing:

Definition: Prompt speaker for responses. (**this excludes rhetorical question**).

> Examples:
>
> "What is your view on that Dr. Foster?" (Questioning.)
>
> "Where are you from?" (Social questioning.)
>
> "Peter!" (Name calling for response.)
>
> "If the majority of people are voting against it, would you still insist?" (Elaborated questioning.)
>
> "Do you agree with this statement?" (Binary question.)

Confronting:

Definition: Response that prompts one speaker to response or engage with another speaker.

> Examples:
>
> "So David pointed out the critical weakness of the system, what is your thought on his critiques, Dr. Foster?"

Information provision behaviour:

Occasionally moderators themselves contribute information for various purposes, including instruction, clarifying information, filling knowledge gap, expressing opinion etc. For the

provided information, we are also interested in the source of the information, and therefore, we have devised three information provision categories (Instruction, Interpretation, Supplement).

Supplement:

Definition: Enrich the conversation by supplementing details or information without immediately changing the target speaker's behavior.

Examples:

"Supposed we live in a world where such behaviour is accepted." (Hypothesis)

"I suggest the best solution is giving everyone equal chances." (Proposal)

"The government announced tax raise from March." (Providing external information)

"I agree with that you said." (Agreement)

"GM means genetic modified." (Providing external knowledge)

"I think people should be given the right to say no!" (Opinion)

Interpretation:

Definition: Clarify, reframe, summarize, paraphrase, or make connection to earlier conversation content.

Examples:

"So basically, what Amy said is that they didn't use the budget efficiently". (Summarisation)

"You said 'I believe GM is harmless,'." (Quote)

"In another word, you don't like their plan.". (Paraphrase)

"My understanding is you don't support this due to moral reason." (Interpretation)

"She does not mean to hurt you but just tell the truth." (Clarify)

"So far, we have Dr. Johnson suggesting…., and Dr. Brown against it because……"(Summarisation)

"Amy saying that to justify the reduction of the wage, but not aiming to induce suffering." (Reframing)

Instruction:

Definition: Explicitly command, influence, halt, or shape the immediate behavior of the recipients.

Examples:

"Please get back to the topic." (Commanding)

"Please stop here, we are running out of time." (Reminding of the rule)

"The red will start now." (Instruction)

"Please mind your choice of words and manner." (social policing)

"Do not intentionally create misconception." (argumentative policing)

"Now is not your term, stop here." (coordinative policing)

Utility:

There are also various other kinds of dialogue acts that are neither contributing information nor seeking information. Since these kinds of dialogue acts are not the focus of our study, we group all the uncovered dialogue acts into a broad category called "Utility". Occasionally, this group of behaviours play an important role to show engagement (e.g. back channelling) and getting attention (e.g. floor grabbing).

All Utility:

Definition: All other unspecified acts.

Examples:

"Thanks, you." (Greeting)

"Sorry." (Apology)

"Okay." (Back channelling)

"Um hm." (Back channelling)

"But, but, but……" (Floor grabbing)

WHO: Target speaker

We are also interested in who the moderator was talking to at the time given the dialogue context. Besides talking to a particular speaker, the moderator can also talk to him/herself, the audience, or everyone.

Examples:

"We are going to start in 10 minutes. The red team will go first." (talking to everyone).

"Paul, what is your thought?" (talking to Paul Helmke)

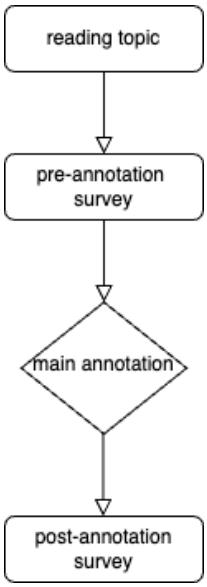
"Cough! Cough!" (Self)

“The guy sitting at the front row. Yes! You!” (talking to Audience)

“This is ‘Intelligence Square’. Welcome back!” (talking to Audience)

Annotation instruction and steps

For every debate annotation task, you will firstly be provided the topic, speakers information, and the debate transcript. The annotation process starts with **reading the debate topic, then complete the pre-annotation survey. After completing the annotation, there are also a few post-annotation questions about the impression of the moderator. Before starting an episode, please make sure you have time to complete the whole episode in the same time block.**



Topic	Abolish the minimum wage
“For” speakers	Russell Roberts, James A. Dorn
“Against” speakers	Karen Kornbluh, Jared Bernstein
Moderator	John Donvan

Label codes for the three facets:

dialogue acts	motivations	target speakers
q (Probing)	I (Informational motive)	1 (Everyone)
w (Confronting)	S (Social motive)	2 (Self)
e (Instruction)	C (Coordinative motive)	3 (Russell Roberts, For)
d (Interpretation)		4 (James A. Dorn, For)

s (Supplement)		5 (Karen Kornbluh, Against)
a (All utilities)		6 (Jared Bernstein, Against)
		7 (Audience)

Debate transcript (blue = For, red = Against, green = Moderator):

21793_0	Russell Roberts	I think part of the problem we have with education right now is that we've subsidized it, which is a lovely idea.
21793_1	Russell Roberts	And as a result, it's pushed up tuition, and it's allowed colleges to raise their prices, their tuition a great deal.
21793_2	Russell Roberts	And as a result, many students have borrowed have a lot of money.
21793_3	Russell Roberts	And as a result, they're in big trouble.
21793_4	Russell Roberts	And especially in a downtime of economic growth when economic growth is so mediocre.
21794_0	John Donvan	Okay.
21794_1	John Donvan	I just-- it's getting a little bit off the minimum wage issue.
21794_2	John Donvan	Fair enough?
21794_3	John Donvan	But that's why I stopped you.
21794_4	John Donvan	Karen Kornbluh to respond.
21795_0	Karen Kornbluh	Yeah, I do think this is really tied to the minimum wage issue because we have to remember that we live in a knowledge economy.
21795_1	Karen Kornbluh	And a country's human capital is what it competes on.
21795_2	Karen Kornbluh	And so what we need to do to be competitive, to have productivity, to have the American dream again, to have people earning high wages and being able to support their families is investing in people's education.
21795_3	Karen Kornbluh	And so we have a big problem in this country in terms of K-12, and we have a big problem in terms of--
21796_0	John Donvan	Okay, for the same reason, Karen--
21797_0	Karen Kornbluh	That's what we should adjust and not the minimum wage.
21798_0	John Donvan	All right.
21798_1	John Donvan	I'm going to step in.
21798_2	John Donvan	But your opponents made the very same argument at the beginning.
21798_3	John Donvan	And I was surprised when you said that you had the moral argument on their side because they were not saying "damn the poor" in any way.
21798_4	John Donvan	They were saying that they feel that the tool, the minimum wage, doesn't function correctly.

21798_5	John Donvan	And I've been wanting to get to that moral argument, but I was hoping somebody in the audience would actually bring it up.

The **red highlighted rows are from the "Against team"**; while the **blue highlighted rows are from the "For team"**, and **the green rows are from the "Moderator". Only the green rows require labels.**

*****Attention: the annotation below is only one of the samples from pilot study to show how the annotation works. The annotation itself is not the golden truth.*****

A whole block of consecutive rows from the same speaker is called a **"response"**. As displayed in the dialogue history, **each response has been segmented into sentences**, **since some response might contain more than one semantic utterance**. For example, in the response 21794, the moderator firstly backchanneled the speaker 3 (Russell Roberts, For), then reminded about getting back to the topic, and then finally called another speaker 5 (Karen Kornbluh, Against) to speak.

The annotation interface will have three columns for the three facets to label like shown below:

Id	Speaker	text	Dialogue act	Motivew	Target speaker
21794_0	John Donvan	Okay.	a	I	3
21794_1	John Donvan	I just-- it's getting a little bit off the minimum wage issue.	e	I	3
21794_2	John Donvan	Fair enough?	q	C, S	3
21794_3	John Donvan	But that's why I stopped you.	s	C	3
21794_4	John Donvan	Karen Kornbluh to respond.	q	I	5

However, **you do not need to label each sentence**. Like the example below, if the dialogue act or the perceived intention of the speaker spans through multiple sentences, you will only need to label the top row.

21798_1	John Donvan	I'm going to step in.	e	C	5
21798_2	John Donvan	But your opponents made the very same argument at the beginning.	i	I	5
21798_3	John Donvan	And I was surprised when you said that you had the moral argument on their side because they were not saying "damn the poor" in any way.			
21798_4	John Donvan	They were saying that they feel that the tool, the minimum wage, doesn't function correctly.			