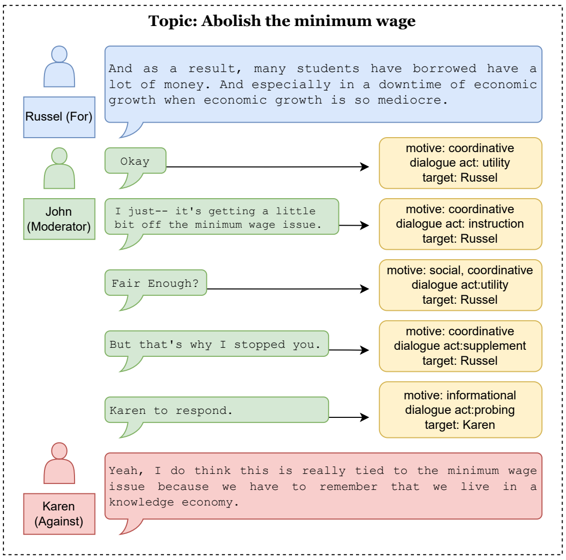
# WHoW: A Cross-domain Approach for Analysing Conversation Moderation

**Ming-Bin Chen** and **Lea Frermann** and **Jey Han Lau**
School of Computing and Information Systems, The University of Melbourne
{mingbin, lfrermann, laujh}@unimelb.edu.au

## Abstract

We propose WHoW, an evaluation framework for analyzing the facilitation strategies of moderators across different domains/scenarios by examining their motives (**W**hy), dialogue acts (**Ho**w) and target speaker (**W**ho). Using this framework, we annotated 5,657 moderation sentences with human judges and 15,494 sentences with GPT-4o from two domains: TV debates and radio panel discussions. Comparative analysis demonstrates the framework's cross-domain generalisability and reveals distinct moderation strategies: debate moderators emphasise coordination and facilitate interaction through questions and instructions, while panel discussion moderators prioritize information provision and actively participate in discussions. Our analytical framework works for different moderation scenarios, enhances our understanding of moderation behaviour through automatic large-scale analysis, and facilitates the development of moderator agents.[1]

## 1 Introduction



And as a result, many students have borrowed have a lot of money. And especially in a downtime of economic growth when economic growth is so mediocre.

Okay

I just-- it's getting a little bit off the minimum wage issue.

Fair Enough?

But that's why I stopped you.

Karen to respond.

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.

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 one of the DEBATE subset conversation, respectively. The peach-colored boxes contain the annotations for the corresponding moderator sentences.

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) mitigation: preventing and policing negative behaviors, such as personal attacks (Gorwa et al., 2020); (2) facilitation: promoting positive and constructive outcomes, such as knowledge generation and consensus building (Vasodavan et al., 2020); and (3) participation: ensuring balance and open participation opportunities for all members (Kim et al., 2020).

Extensive research has focused on content moderation analysis and automation in online spaces, primarily aimed at mitigating negative behaviors and intervening through asychronous actions such as post deletion (Gorwa et al., 2020; Park et al., 2021; Wulczyn et al., 2017; Falk et al., 2024). However, there are few studies that examine how moderators facilitate positive outcomes and balance participation through conversational engagement;

[1] Our code, dataset are available at `https://github.com/mrknight21/conversation_moderation_analysis`.

*Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2091–2126
April 29 - May 4, 2025 ©2025 Association for Computational Linguistics

our study seeks to address this.

We introduce WHoW: an analytical framework that breaks down the moderation decision-making process into three key components: motives (**W**hy), dialogue acts (**Ho**w), and target speaker (**W**ho). Using this framework, we analyzed transcripts of conversation moderation in two distinct domains: Intelligent Squared TV Debate (DEBATE) and Roundtable Radio Panel (PANEL). We began by annotating 50 episodes of transcripts with human annotators, which are then used to create and evaluate prompts for GPT-4o (OpenAI, 2024) for automatic annotation. We then used GPT-4o to automatically annotate more data and compared the moderation strategies in facilitating and balancing participation in these two domains. Our findings reveal distinct moderation strategies and have the potential to support moderator training/assessment and facilitate the future development of moderator agents.

To summarise, our key contributions are:

1. We develop an analytical framework that characterizes conversational moderation across different scenarios using three dimensions: motives (Why), dialogue acts (How), and target speaker (Who);

2. Based on the framework, we annotated moderated multi-party conversations in two domains: TV debates and radio panel discussions. Our dataset comprises a total of 5,657 human-annotated sentences and model-annotated 15,494 sentences (GPT-4o).

3. By analyzing these two conversational domains—debates and panel discussions—we demonstrate the framework's cross-domain generalizability and identify distinct moderation strategies. Debate moderators focus on coordination, facilitating interactions through follow-up and confrontational questions, as well as instructions. In contrast, panel discussion moderators actively engage in and contribute to the topic themselves, while being less involved in fostering interactions between speakers.

## 2 Related Work

Conversation 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; Wei et al., 2023) must account for the conversation context, group dynamics, and the balance of participation. Depending on the scenario, moderators fulfill various functional roles, such as inviting for contribution, providing background information, facilitating topic transitions, and posing questions to guide discussions and maintain their quality (Wright, 2009; Park et al., 2012; Mao et al., 2024; Schroeder et al., 2024). 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 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), online public issue discussions and debates (Wright, 2009; Edwards, 2002), and 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, the research often relies on methodologies such as interviews or case studies, which are not reusable for further analysis or automation (Forester, 2006). As of the time this study's manual annotation was conducted, the only annotated dataset currently available consists of just 300 comments (Park et al., 2012)[2]. Furthermore, some studies treat moderation as a reactive intervention to participant comments and structure the 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) and don't generalise to other domains.

## 3 The WHoW Conversational Moderation Analytic Framework

We design an analytical framework that: (1) is grounded in the existing literature (Wright, 2009; Park et al., 2012; Vasodavan et al., 2020; Lim et al., 2011); (2) captures the multifaceted nature of conversational moderation; and (3) generalizes across different domains. Our framework (Table 1), inspired by existing multi-party agent work (Wei et al., 2023; Mao et al., 2024), is structured around three core dimensions: motives (Why), dialogue acts (How) and target speakers (Who). In addition to dialogue acts, which are widely employed to study dialogue patterns (Shriberg et al., 2004), we incorporate the motive dimension to provide insights into the *reason* the moderator intervenes given a particular scenario or context (Yeomans et al., 2022). Furthermore, we introduce 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.

Table 1 shows the label definitions under the three dimensions. To derive our labels and understand their compatibility with existing protocols, we categorize all moderation-related typologies identified in Section 2 into motives and dialogue acts, as detailed in Appendix Table 9. Since moderator responses can be lengthy, and may serve multiple goals (i.e., correspond to multiple labels), we first break them into individual sentences and then label each sentence across the three dimensions using the definitions provided above. We elaborate on these three dimensions in the following sections.

### 3.1 Motives: Why does the moderator intervene?

The "Why" component examines the motivations behind a moderator's interventions in conversations. Existing protocols distinguish socially motivated speech — such as "affective strategy" (Hsieh and Tsai, 2012) and "social functions" (Park et al., 2012) — from argument-driven speech. This pattern aligns with the conversational circumplex framework, which categorizes conversational goals along informational and relational dimensions (Yeomans et al., 2022). Furthermore, in facilitated group debates like Intelligent Squared Debate (Zhang et al., 2016) moderator interventions can be motivated by meeting rules, such as adherence to time limits. Consequently, we propose three motives driving moderation behaviors: informational, social, and coordinative (Table 1, top), which align with the facilitation types described by Lim et al. (2011) and accommodate the hybrid-motive nature of the moderator role. As previous studies (Yeomans et al., 2022) and our pilot studies show that a single speech can convey multiple motives, we treat this annotation as a multi-label task (e.g., a moderator sentence may have both social and coordinative motives).

### 3.2 Dialogue Acts: How does the moderator intervene?

The "How" component analyzes the dialogue acts, or the immediate functions of a moderator's interventions. By examining the sequential patterns of these acts, we gain insights into the strategies moderators use to realize their motives. The initial set of dialogue acts is derived from the five fundamental labels of the MRDA corpus (Shriberg et al., 2004), which was developed for annotating multi-party meetings: "Question', "Statement", "BackChannel", "Disruption", and "FloorGrabber". The two major labels, "Question", and "Statement", indicate the functions of information elicitation and information provision respectively. These two major labels are instrumental in distinguishing the moderator's functional role as either a "Interviewer" or an "Contributor" respectively (McLafferty, 2004). The remaining MRDA labels, along with other unspecified acts, such as greeting, are grouped into the 'Utility' category, as they do not directly contribute

[2] In concurrent work, Schroeder et al. (2024) published a substantial facilitative conversation corpus.

| **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 "unknknown".

to information exchange."

We further categorise the two major labels into sub-labels to capture the nuanced characteristics of moderators' interventions. For "Question", we distinguish two types of information elicitation interventions to capture whether the moderator seeks to acquire information through direct prompts (**Probing**) or by encouraging interaction among participants (**Confronting**). Turning to "Statement", we distinguish between interjections that change a participant's behavior (e.g. command to stop; **Instruction**), refer back to prior discussion (e.g. summarization; **Interpretation**), and provide additional information or opinions (**Supplement**) (Park et al., 2012; Wright, 2009).

The detailed definitions of these fine-grained labels are included in Table 1 (middle). Appendix Table 10 presents example sentences that intersect between the motives and dialogue acts dimensions. We treat dialogue acts as mutually exclusive and formalize it as a multi-class classification task.

### 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 allows for an assessment of the intended shifts in participation and the moderator's initiatives during the discussion.

We approach the annotation of this dimension as a multi-class classification task, with labels corresponding to speakers. To accommodate different contexts, we also introduce general labels such as "everyone" (including audience), "unknown", and "all speakers". For the TV debates domain specifically we introduce 3 additional labels "audience", "against team", and "support team". While our framework is designed to be cross-domain, we note that its labels or categories are customizable depending on the domain.

## 4 Dataset and Human Annotation

### 4.1 Datasets

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

|  | DEBATE |  |  | PANEL |  |
|---|---|---|---|---|---|
|  | **Test** | **Dev** | **Train** | **Test** | **Train** |
| Episodes | 19 | 11 | 78 | 20 | 68 |
| Speakers / episode Mean | 4.63 | 4.55 | 4.62 | 3.450 | 4.47 |
| 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 DEBATE and PANEL. M denotes Moderator; share the proportion of words uttered by the moderator; and turn the full utterance (which contains multiple sentences).

|  | **DA** | **IM** | **CM** | **SM** | **TS** |
|---|---|---|---|---|---|
| DEBATE | 0.49 | 0.43 | 0.37 | 0.41 | 0.72 |
| PANEL | 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 DEBATE and PANEL.

perts 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, where the majority of the moderated interactions occur.[3] We randomly split the episodes into 11 for development, 19 for testing, and 78 for training.

To validate the generalizability of our framework across scenarios, we also include a second dataset from a subset of The NPR Interview Corpus (Majumder et al., 2020) (henceforth PANEL). We specifically select 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 (unlike DEBATE). This selection yielded 88 episodes, from which we randomly sampled 20 episodes to create a test set. Table 2 presents some descriptive statistics of the two datasets.

### 4.2 Human annotation process

We recruited annotators to label each sentence of the moderator's utterance based on the WHoW framework, as illustrated in Figure 1.[4] We recruited five annotators in total, all proficient or native English speakers and students of either linguistics or NLP, and they were paid 36 USD/hour. The annotators manually annotated the development and test sets of DEBATE and the test set of PANEL. Annotators received the definitions of labels as outlined in Section 3 and Table 1. To facilitate the dialogue act annotation and increase agreement, we developed a decision tree flowchart (see Appendix Figure 3). 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. More details of annotation material and interface are provided in Appendix Section F.

Each sentence in the moderators' utterance was annotated for the presence of the three motives, one identified dialogue act, and the target speaker(s). Each episode was annotated by at least two annotators. The final ground truth were aggregated using majority vote; in cases of evenly divided annotation votes, the first author did the tie-breaking. Inter-annotator agreement (Krippendorff's alpha) is presented in Table 3. PANEL generally has higher agreement and the overall agreement ranges from moderate to good and these numbers are consistent with previous studies that involve complex and subjective judgements (Falk et al., 2024). A detailed analysis of disagreements is provided in Appendix section C.

## 5 Automatic Annotation

Manual labeling is time-consuming and extensive. A practical and generalizable framework for large-scale exploration requires an automatic labeling framework. To this end, we leverage GPT-4o (OpenAI, 2024) for automatic annotation. We optimized the prompts using the development set from DEBATE (see Appendix section B. for more details on the prompt design). Our single-task setting ("ST") frames the annotation as five independent classification tasks: two multi-class classifications for dialogue acts ("DA") and target speakers ("TS"), and three binary classifications for motive labels ("IM", "CM", "SM"). In addition to ST, we also developed an alternative approach to perform all

[3] In addition, the corpus provides information on each speaker's role (moderator, team member, audience member) and metadata, such as short bios and audience voting results before and after the debate.

[4] The development of the annotation schema started with two rounds of pilot studies involving the paper authors and NLP PhD students for testing the preliminary definitions of the framework with one episode from each dataset. Feedback from the pilot resulted in framing the motive labels as a multi-label task and reducing the dialogue act classes from eight to six with refined definitions.

| Model | DA | IM | CM | SM | TS |
|---|---|---|---|---|---|
| Random(DEBATE) | 0.153 | 0.492 | 0.508 | 0.405 | 0.057 |
| MT(DEBATE) | 0.485 | **0.761** | **0.711** | **0.767** | 0.497 |
| ST(DEBATE) | **0.515** | 0.7287 | 0.686 | 0.668 | **0.525** |
| Random(PANEL) | 0.115 | 0.490 | 0.482 | 0.387 | 0.096 |
| MT(PANEL) | **0.504** | 0.726 | **0.732** | **0.754** | **0.467** |
| ST(PANEL) | 0.492 | **0.747** | 0.639 | 0.635 | 0.464 |

Table 4: Macro-F1 comparing GPT-4o 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 are derived from five random simulations.

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

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

| | prob | conf | inst | inte | supp | util | $p(m)$ |
|---|---|---|---|---|---|---|---|
| DEBATE | | | | | | | |
| IM | **0.41** | $\underline{0.23}$* | 0.04* | 0.11* | 0.20 | 0.01 | 0.39 |
| CM | $\underline{0.15}$* | 0.10* | **0.54*** | 0.02 | 0.09 | 0.10 | 0.66* |
| SM | 0.08 | 0.02 | 0.10* | 0.02 | $\underline{0.14}$ | **0.65** | 0.12 |
| $p(d)$ | 0.22 | 0.11* | 0.36* | 0.05* | 0.12 | 0.14* | |
| PANEL | | | | | | | |
| IM | $\underline{0.42}$ | 0.03 | 0.01 | 0.03 | **0.51*** | 0.01 | 0.72* |
| CM | 0.06 | 0.02 | **0.42** | 0.01 | $\underline{0.33}$* | 0.17* | 0.25 |
| SM | 0.05 | 0.01 | 0.02 | 0.01 | $\underline{0.28}$ | **0.63** | 0.16* |
| $p(d)$ | 0.30* | 0.03 | 0.10 | 0.02 | 0.41* | 0.13 | |

Table 6: Conditional probabilities of dialogue acts ($d$; columns) given motives ($m$; rows), with marginal probabilities averaged across episodes for the two scenarios—DEBATE (top) and PANEL (bottom). The most likely dialogue act per motive is highlighted in bold, and the second most likely is underlined. * indicates a significantly larger $p(d|m)$ in one data set compared to the other (t-test; p<= 0.05). On average, a moderator speaks 151 sentences per DEBATE episode and 61 per PANEL episode.

tasks jointly with one single prompt (multi-task or "MT"). We present macro-F1 and agreement results of ST and MT with human test set annotations in Table 4 and Table 5 respectively. Overall, the results are encouraging and demonstrate that GPT-4o is a viable method for automatic annotation, particularly given the tasks' high level of subjectivity and complexity (Falk et al. (2024); Appendix Section C). Error analysis (Appendix Section E) reveals that most mis-classifications arise from subjective interpretations, context dependency, or ambiguity in extremely long or short sentences. As the multi-tasking approach (MT) has higher average Macro-F1 (0.64 vs. 0.61) and agreement (0.51 vs. 0.46) across tasks and datasets, we used this approach for automatic annotation, and ran it on the training sets of DEBATE and PANEL.[5]

## 6 Analysis

We now analyze which dialogue acts are used to facilitate discussion and encourage participation. To demonstrate the generalizability of the automatic framework, this analysis draws on our full dataset, including development, test, and train sets, annotated using GPT-4o with the multi-task approach. In Appendix Section D, we compare GPT-4o labels against human label distributions on the development and test set, showing that they are overall consistent, with the exception of the "instruction" and "supplement" acts, with only minor variations in magnitude. Specifically, we examine how speaker rotation is facilitated and how the three motives are addressed across the two domains in the dataset.

### 6.1 Motives and Dialogue Acts

Table 6 presents the probabilities that motive $m$ (rows) is realized by dialogue act $d$ (columns), $p(d|m)$, as well as all dialogue acts and motives labels' marginal probabilities. There is a distinct difference in relative motive frequencies between the two domains. DEBATE moderation is dominated by a coordinative motive (66%) followed by informational (39%). In contrast, informational motives are the most frequent in PANEL moderation (72%). For relative dialogue act frequencies, DEBATE moderators mostly focus on providing instructions (36%), while PANEL moderators tend to supply information (41%). Probing is the second most common dialogue act in both corpora.

Turning to the conditional probabilities, strategically, DEBATE moderators achieve **informational**

[5] We also experimented with using the automatically annotated training sets to fine-tune smaller supervised language models, such as Longformer (Beltagy et al., 2020), and found it works quite well; see Appendix Section B.1.

**motives** (IM) by actively facilitating participant contributions through methods such as probing (0.41) and confronting (0.23), along with notable uses of interpretation (0.12) and supplementing (0.19) information. IM in PANEL, on the other hand, is characterized by moderators delivering information themselves (0.51) and engaging participants through probing (0.41). The minimal use of confrontation (0.03) and interpretation (0.01) in PANEL indicates relatively few attempts to foster interaction and engagement between non-moderator participants.

Conversely, DEBATE moderators more frequently prompt participants to respond to one another (confronting) and engage with earlier discussion content (interpretation). Overall, for IM DEBATE moderators' interventions are more diverse and leading to interaction between participants compared to those in PANEL.

**Coordination motives** (CM) in both domains primarily rely on instructions (0.54 in DEBATE and 0.42 in PANEL). However, DEBATE moderators are more likely to coordinate through probing (0.15), maintaining dialogue engagement by asking participants about their preferences for rotation and participation. PANEL moderators coordinate by providing supplementary information (0.33), e.g. by explaining rules.

While PANEL has a significantly higher proportion of moderator interventions driven by **Social Motives** (SM) compared to DEBATE, there is no notable difference in the dialogue acts used. Both settings primarily utilize utility acts (0.65 in DEBATE, 0.62 in PANEL), such as greetings, along with some social/personal information sharing (supplement) to fulfill their social motives. Although our observations can be partially explained by the respective rules of the discussion programs, they highlight different high-level strategies to facilitate a constructive discussion.

### 6.2 Balancing Speaker Participation

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 and speaker rotation.

Given an episode of a conversation consisting of $n$ turns between a moderator and participants, we denote the speaker identities (e.g. moderator or name of a participant) as $[p_0, p_1, \dots, p_n]$. Note that a turn here denotes the full utterance (which can have multiple sentences) by a speaker.

| DEBATE | | | |
|---|---|---|---|
| | moderation | continuation | rotation |
| moderation | – | 0.52 | 0.48 |
| continuation | 0.78 | – | 0.22 |
| rotation | 0.47 | – | 0.53 |
| PANEL | | | |
| moderation | – | 0.35 | 0.65 |
| continuation | 0.80 | – | 0.20 |
| rotation | 0.60 | – | 0.40 |

Table 7: Transition probabilities between moderator interventions and speaker rotation / continuation. Note: the transition from 'rotation' to 'rotation' represents instances of participant-driven rotation without moderator intervention. '–' indicates that transitions are not possible.

To understand the rotation pattern (i.e. how the dialogue transition from one speaker to another), we simplify the speaker status for each turn ($s_t$) as follows:

$$s_t = \begin{cases} \textit{moderation}, & \text{if } p_t = \text{moderator} \\ \textit{continuation}, & \text{if } p_t \neq \text{moderator } \& \\ & \quad p_t = p_{t'} \\ \textit{rotation}, & \text{if } p_t \neq \text{moderator } \& \\ & \quad p_t \neq p_{t'} \end{cases}$$

where $t'$ is the last non-moderator turn before $t$.

By converting the conversation sequence into three states—moderation, continuation, and rotation—we derive a transition probability matrix ($P(s_{t+1}|s_t)$), as shown in Table 7. The table reveals several key patterns: both DEBATE and PANEL moderators are more likely to intervene when a speaker has continued for more than one exchange (0.78 and 0.80). However, DEBATE moderators (0.52) exhibit a higher tendency than PANEL moderators (0.35) to continue the conversation with the same participant; or another interpretation is that PANEL moderator intervention has a higher tendency to lead to speaker *rotation*. Additionally, there are more participant-driven rotations ($rotation \rightarrow rotation$) in the DEBATE dataset (0.53) compared to the PANEL dataset (0.40), indicating a higher level of *independent* interaction among participants in DEBATE.

We next use the dialogue act (DAs) to further investigate how rotations are facilitated. For each

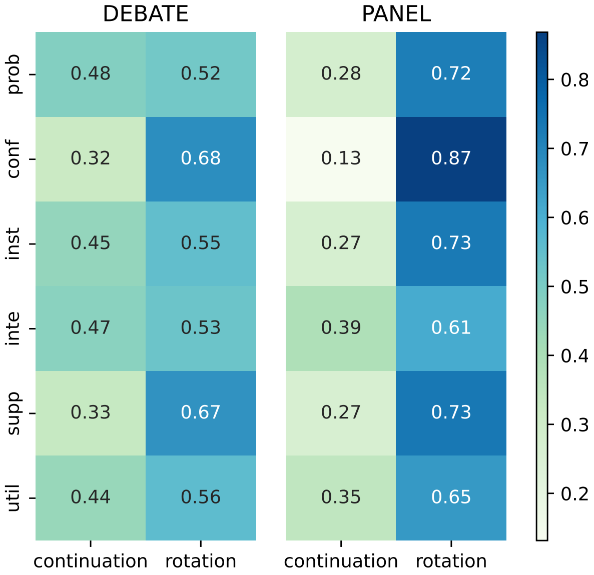


Figure 2: Probabilities of participants' rotation statuses following different moderation dialogue acts.

| | **Pro-activity** | **Interactivity** | **Specificity** |
|---|---|---|---|
| DEBATE | 0.59 | 0.73 | 0.63 |
| PANEL | 0.61 | 0.75 | 0.85 |

Table 8: Proportion of moderator sentences that are pro-active ($target\_speaker \neq last\_speaker$), interactive ($target\_speaker = next\_speaker$), and specific (targeting an individual).

moderator turn $t$ (i.e. $p_t$ = moderator) and denoting the dialogue acts as $d_t$, we compute $P(s_{t+1}|d_t)$ and present the results in Figure 2.[6] We see that moderator intervention in PANEL tends to lead to speaker *rotation* across all dialogue acts. Most dialogue acts in DEBATE, however, lead to both *continuation* and *rotation* almost equally; the only exceptions are confronting and supplementary. This is perhaps not surprising, as confronting questions are designed to explicitly prompt one speaker to respond to another speaker's statement.

### 6.3 Moderators' selection of target speakers

Moderators may exhibit different interaction styles with participants depending on the context, in terms of pro-activity (how often the moderator actively initiates conversations), interactivity (how likely participants respond to the moderator), and specificity (how often the moderator addresses specific individuals rather than the group as a whole) (Wagner et al., 2022). For instance, in a highly scripted setting, a moderator may act primarily as an assistant, responding to participant queries and broadcasting reminders about time and rules, with no expectation of responses—showing low levels of pro-activity, interactivity, and specificity. In contrast, in a more dynamic setting, a moderator might initiate conversations by asking questions tailored to individual speakers.

By analyzing whether the moderator's target speaker aligns with the speakers preceding and following their intervention, we can infer the moderator's interaction style. For each moderator turn $t$ (i.e. $p_t$ = moderator), we denote the set of target speakers as $r_t$ (a moderator turn can have multiple sentences and hence multiple target speakers) and compute pro-activity, interactivity and specificity as follows:

$$\text{pro-activity} = \frac{\#(p_{t-1} \notin r_t)}{M}$$
$$\text{interactivity} = \frac{\#(p_{t+1} \in r_t)}{M}$$
$$\text{specificity} = \frac{\#(r_t \subset S)}{M}$$

where $M$ is the total number of moderator turns and $S$ the set of unique participants in the conversation.

Table 8 indicates that moderators in both domains demonstrate high levels of pro-activity and interactivity, suggesting that they frequently initiate interactions with participants. However, PANEL moderators exhibit higher levels of specificity compared to DEBATE moderators, indicating a greater tendency to address specific individuals rather than the group as a whole. This suggests that PANEL moderators are more likely to tailor their interventions to particular participants, fostering more targeted and personalized interactions.

## 7 Conclusion

We present WHoW, an analytical framework that characterizes conversational moderation across domains. WHoW breaks down the complexity of moderation decision-making into three key components: why the moderator intervenes (motives), how they intervene (dialogue acts), and to whom they direct the intervention (target speakers). Using this framework, we annotated moderation utterances in two distinct scenarios: Intelligent Squared Debate Corpus (DEBATE) and RoundTable Radio

[6] Each moderator turn may have one or more dialogue acts (since there can be multiple sentences for a turn). All dialogue acts contribute to the transition matrix counts. For example, if $d_t$ = [prob, inst] and $s_{t+1} = rotation$, we have 2 transitions: prob → $rotation$ and inst → $rotation$.

Panel Discussion (PANEL). We showed that GPT-4o can effectively automate the labelling process. In total, our dataset has 5,657 human-annotated and 15,494 GPT-4o annotated moderation sentences, which is an order of magnitude larger than existing datasets (Park et al., 2012).

Our analysis demonstrates the framework's effectiveness in differentiating intervention strategies and styles across the two scenarios. In DEBATE, moderators are primarily coordination-motivated, serving functional roles as interviewers and instructors, while occasionally facilitating interaction between non-moderator speakers. In contrast, PANEL moderators are more information-oriented, acting as both contributors and interviewers, as they often participate in the discussion topics. While they seek information from the speakers and balance turn-taking, they promote less direct interaction between non-moderator speakers.

Our framework can serve as an exploratory tool or foundational skeleton for domain-specific adaptation and expansion in moderation analysis or moderator agent development. Using raw transcripts, users can initially categorize the moderator's speech into the twelve categories across the motive and dialogue act dimensions, then refine these labels based on the specific domain context. For example, in a mental health support setting, "social interpretation" could be expanded into more specific categories like "emotion interpretation". Although the current dataset may not be large enough for fine-tuning supervised models, it serves as a valuable resource for in-context few-shot learning. This means it provides practical examples that help develop models capable of predicting or recommending intuitive moderation interventions based on our framework.

Future studies should encompass a broader range of moderation scenarios, such as group counseling (Kissil, 2016) and second language group conversations (Gao et al., 2024). Additionally, the proposed analytic framework could be expanded to support the generation of conversational moderation strategies by sequentially predicting the three key components. Another important direction is the development of evaluation metrics to assess the effects and potential biases of moderation interventions (Spada and Vreeland, 2013), enabling deeper insights into the impact and fairness of moderation practices. Finally, a broader goal for future work could involve synthesizing the results into "moderator prototype strategies" — a schema with multiple axes capturing distinct moderation styles. As new scenarios are explored, these prototypes may evolve, offering a richer understanding of diverse moderation approaches.

## 8 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 identifiable personal information of the annotators has been removed from the datasets. Since the annotations are based on publicly available datasets (Zhang et al., 2016; Majumder et al., 2020), there are no confidentiality concerns regarding the speakers' privacy or personal information. The annotation protocol and material were approved by the University of Melbourne research ethics committee with the reference code- 2023-28400-47354-1.

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.

## Acknowledgments

I would like to express my gratitude to Rena Gao, Rui Xing, Gisela Vallejo, and Miao Li, my wonderful colleagues who contributed significantly during the pilot study stage. I also wish to thank the anonymous workers who participated in the annotation process for this project. Additionally, I am grateful to Alex Goddard for sharing inspiring literature and insights that greatly informed this research.

## 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.

Soomin Kim, Jinsu Eun, Changhoon Oh, Bongwon Suh, and Joonhwan Lee. 2020. Bot in the bunch: Facilitating group chat discussion by improving efficiency and participation with a chatbot. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–13.

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.

Sze Chung Raymond Lim, Wing Sum Cheung, and Khe Foon Hew. 2011. Critical thinking in asynchronous online discussion: An investigation of student facilitation techniques. *New Horizons in Education*, 59(1):52–65.

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.

Manqing Mao, Paishun Ting, Yijian Xiang, Mingyang Xu, Julia Chen, and Jianzhe Lin. 2024. Multi-user chat assistant (muca): a framework using llms to facilitate group conversations. *arXiv preprint arXiv:2401.04883*.

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.

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

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.

Nicolas Wagner, Matthias Kraus, Tibor Tonn, and Wolfgang Minker. 2022. Comparing moderation strategies in group chats with multi-user chatbots. In *Proceedings of the 4th Conference on Conversational User Interfaces*, pages 1–4.

Jimmy Wei, Kurt Shuster, Arthur Szlam, Jason Weston, Jack Urbanek, and Mojtaba Komeili. 2023. Multi-party chat: Conversational agents in group settings with humans and models. *arXiv preprint arXiv:2304.13835*.

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*.

Ming Zhong, Yang Liu, Yichong Xu, Chenguang Zhu, and Michael Zeng. 2022. Dialoglm: Pre-trained model for long dialogue understanding and summarization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11765–11773.

# 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), invite feedback or comments(5), questioning clarifications probe viewpoints(5), open invitation (6), specific intivation to participate (6), follow-up question (6) | empathetic exploration(4), participation encouragement (7) | coordinative enquiry* | 0: Park et al. (2012),<br>1: Vasodavan et al. (2020),<br>2: Hsieh and Tsai (2012),<br>3: Wright (2009),<br>4: Sharma et al. (2020),<br>5: Lim et al. (2011),<br>6: Schroeder et al. (2024),<br>7: Mao et al. (2024),<br>*: 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), make connections (6) | conflict resolver (3), conflict resolution (7) | 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), establish new threads/directions (5) | |
| Inte | correcting misstatements or clarifying (0), summarizaing discussion (1), highlight contribution (1), archiving information (1), summarizer of debates (3), summarize salient points (5), initiative summarization (7) | 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), ' ybrarian' (3), expressing agreements(5), challenge others' viewpoints (5), make connections with supporting research (5), providing opinions/explanation (5), express agreements or affirmation (6), model examples (6), in-context chime-in (7) | informal talk (1), adding personal experience/opinion (1), welcomer (3), empathetic reaction(4), direct chatting (7) | 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), acknowledgement or showing appreciation (5), express appreciation (6) | keep silent (7), floor grabbing* | |

Table 9: This table presents a collection of literature with taxonomies for moderation/facilitation, mapping their classifications across the dialogue acts and motives dimensions of our framework.

| DAs | IM | CM | SM |
|---|---|---|---|
| Prob | Can you take that on? (prompting)<br>As long as the political spectrum is covered overall, what's wrong with that? (follow up question)<br>Siva? (name calling prompt) | Which of you would like to go first? (preference inquiry)<br>Did this gentleman come down yet? (coordinative question)<br>It's working, right? (question managing environment) | Is that a relief to you or– (asking feeling)<br>Could you tell us your name, please? (social question)<br>Do you have eyeglasses? (humour question) |
| Conf | That landed pretty well I think, so can you respond to that? (counter confronting)<br>On this side, do you want to respond, or do you agree? (consensus confronting)<br>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)<br>Response from the other side, or do you want to pass? (coordinative confronting)<br>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)<br>I'd love to hear your answer to that question, so go for it. (confronting with affective appeal)<br>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)<br>Relate that point to this motion. (back to topic)<br>I want to stay on the merits of the Obama plan. (manage topic) | Remember, about 30 seconds is what you'll get. (time control)<br>Can you go up three steps, please, and turn right? (coordinating instruction)<br>I'll be right back after this message. (program management) | Do not be afraid. (emotion instruction)<br>Those who agree, just a round of applause to that. (pro-social instruction)<br>–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. (summarization)<br>Their point is that it would be a bad thing. (simplification)<br>But that would be the question of mobility. (reframe) | That was an ambiguous signal. (situation interpretation)<br>You're pointing to Lawrence Korb.(preference interpretation)<br>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)<br>And it's a little bit insulting almost to say (toxicity interpretation)<br>—honestly, I don't think that was an—a personal attack— (toxicity interpretation) |
| Supp | I agree that it is.(agreement)<br>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)<br>They had never paid any attention whatsoever to Africa. (share opinion) | Fifty-one of you voted against the motion. (vote reporting)<br>And the mic's coming down to you. (describe situation)<br>Round two is where the debaters address each other directly (rule explanation) | You have a colorful sleeve. (social chit-chat)<br>I hate to reward it but I'm going to. (encouragement)<br>And I think all of us probably share a sense that we want things to improve. (state common feeling) |
| Util | Fair question. (acknowledgement)<br>Right (acknowledgement)<br>So the– (floor grabbing) | All right. (backchanneling)<br>Actually, I– (floor grabbing)<br>Well—(floor grabbing) | Thank you Evgeny Morozov. (thanks)<br>I'm sorry. (apology)<br>Hi. (greeting) |

Table 10: This table presents a collection of exemplar sentences at the intersection of the motives and dialogue acts dimensions.

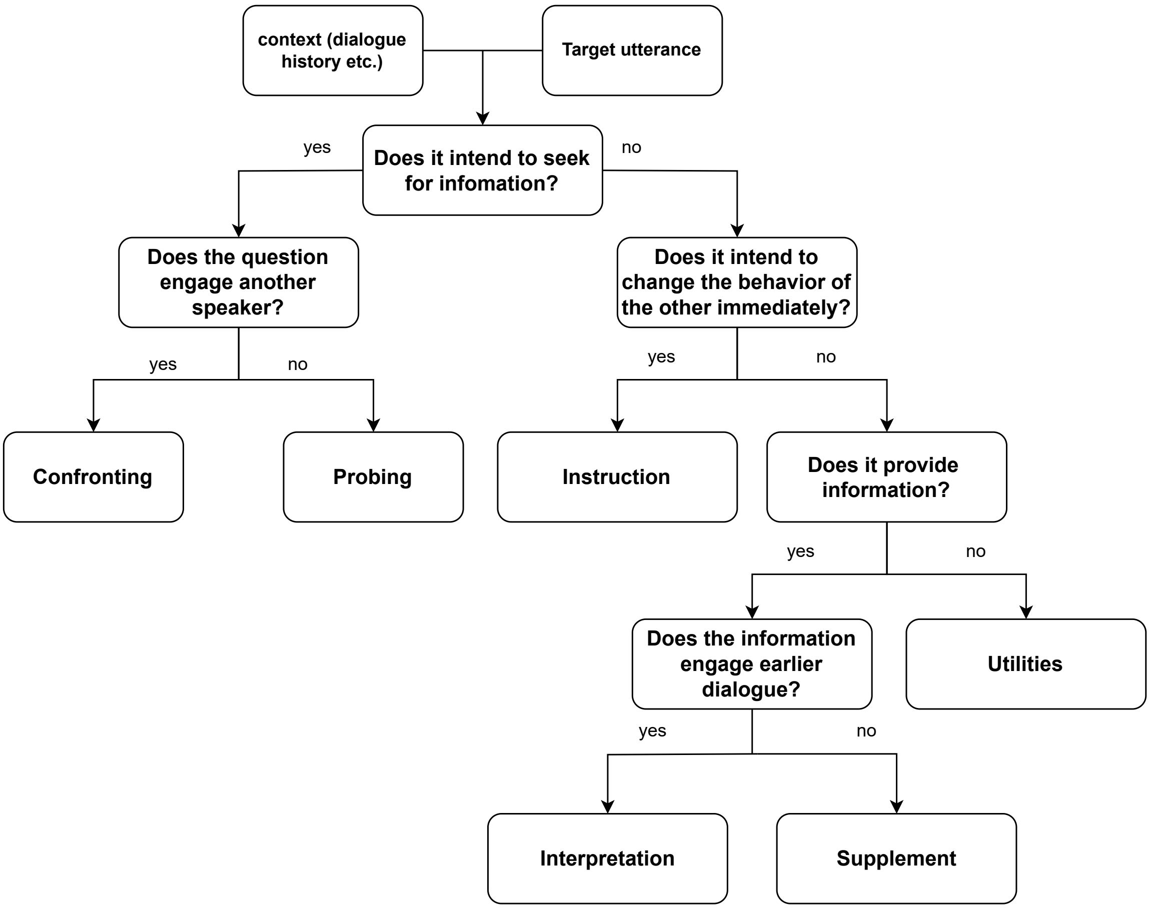

context (dialogue history etc.)
Target utterance
Does it intend to seek for infomation?
yes
no
Does the question engage another speaker?
Does it intend to change the behavior of the other immediately?
yes
no
yes
no
Confronting
Probing
Instruction
Does it provide information?
yes
no
Does the information engage earlier dialogue?
Utilities
yes
no
Interpretation
Supplement

Figure 3: The decision tree used by annotators to resolve ambiguous sentences that may involve multiple dialogue acts.

## B Prompt Engineering

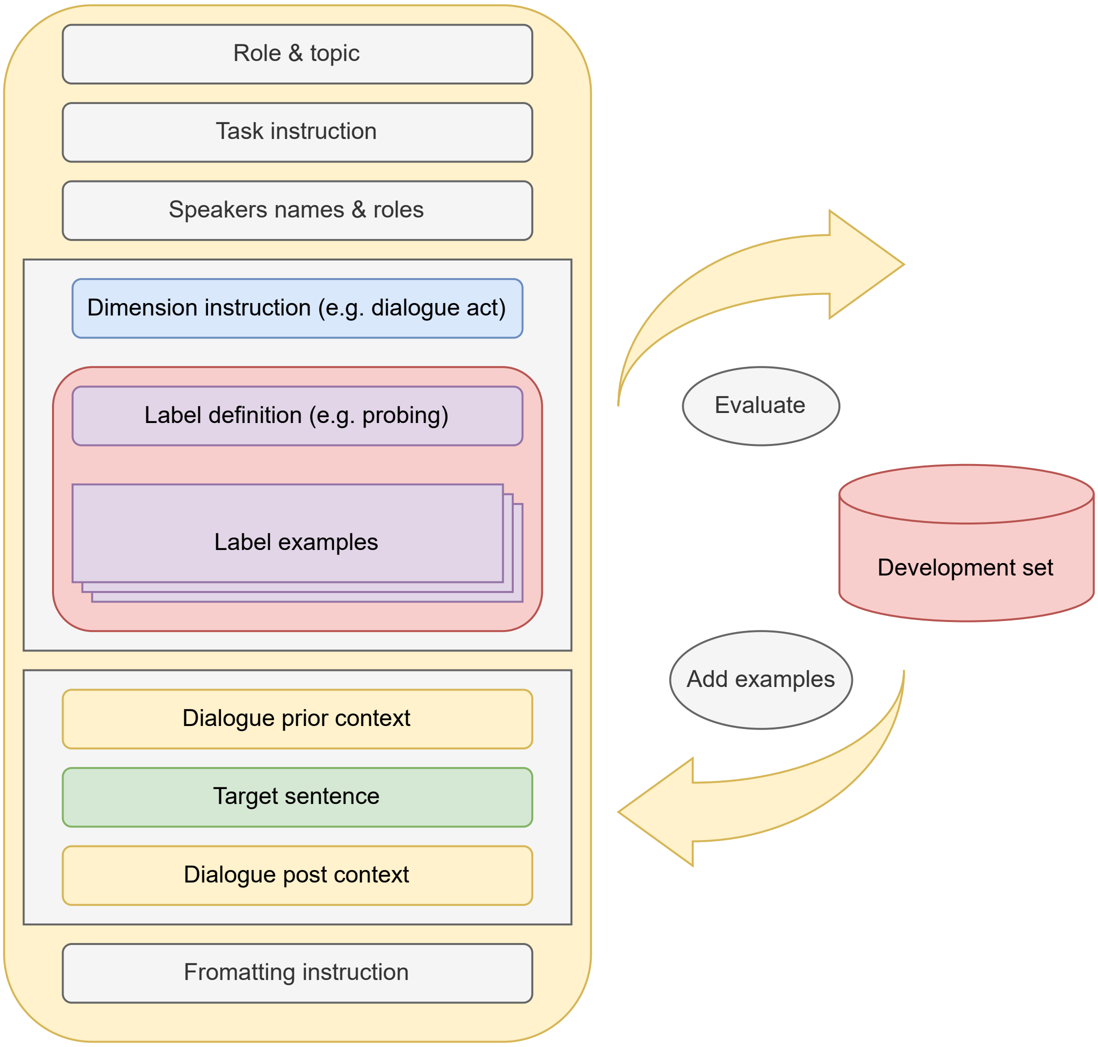


Figure 4: Prompt structure and development cycle

Our prompt design, as illustrated in Figure 4, 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 11 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 12 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:<br><br>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.<br><br>Evgeny Morozov (against): That was your publicist.<br><br>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–<br><br>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.<br><br>Eli Pariser (for): Right, and the question is, can you trust them?<br><br>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:<br><br>John Donvan (mod): That landed pretty well I think, so can you respond to that? |
| Dialogue post context | Dialogue context after the target sentence:<br><br>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.<br><br>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 11: 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:<br>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.)<br>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.)<br>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:<br>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.)<br>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?"<br>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)<br>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”. (Summarization) “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. . . . . . ”(Summarization) “Amy saying that to justify the reduction of the wage, but not aiming to induce suffering.” (Reframing)<br>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}<br>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 12: An example of a multi-task prompt. Here we only demonstrate the components that are different from the single-task prompt.

### B.1 Supervised model training and comparison

| Model | DA | IM | CM | SM | TS |
|---|---|---|---|---|---|
| Random (DEBATE) | 0.153 | 0.492 | 0.508 | 0.405 | 0.057 |
| GPT-4o-MT(DEBATE) | 0.485 | 0.761 | 0.711 | 0.767 | 0.497 |
| GPT-4o-ST(DEBATE) | **0.515** | 0.729 | 0.686 | 0.668 | **0.525** |
| longformer-MT(DEBATE) | 0.494 | 0.764 | 0.719 | **0.784** | 0.246 |
| longformer-ST(DEBATE) | 0.493 | **0.772** | 0.726 | 0.694 | 0.299 |
| DialogLMLED-MT(DEBATE) | 0.489 | 0.760 | **0.760** | 0.714 | 0.147 |
| Random(PANEL) | 0.115 | 0.490 | 0.482 | 0.387 | 0.096 |
| GPT-4o-MT(PANEL) | **0.504** | 0.726 | 0.732 | 0.754 | **0.467** |
| GPT-4o-ST(PANEL) | 0.492 | 0.747 | 0.639 | 0.635 | 0.464 |
| longformer-MT(PANEL) | 0.414 | 0.753 | **0.774** | 0.731 | 0.196 |
| longformer-ST(PANEL) | 0.417 | 0.757 | 0.759 | 0.729 | 0.225 |
| DialogLMLED-MT(PANEL) | 0.389 | **0.764** | 0.751 | **0.768** | 0.132 |

Table 13: 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.

To further explore training smaller language models for motive and dialogue act classification, we fine-tuned the Hugging Face pre-trained Longformer model (allenai/longformer-base-4096) (Beltagy et al., 2020). The input sequence included the discussion topic; a list of speaker options comprising all speaker names along with "unknown," "everyone," "audience," and "all speakers"; and, for the DEBATE subset, additional options "against team" and "support team." We also incorporated the five utterances preceding and the two utterances following the target sentence, with a maximum input length of 3,072 tokens. The model was trained for three epochs over three hours using a learning rate of 2e-5 with the AdamW optimizer (weight decay = 0.01) and a batch size of 8 on an A100 GPU via the Spartan cluster.

We compared both single-task and multi-task variants of the Longformer, employing individual and combined loss functions, respectively. 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. Additionally, recognizing that the original Longformer models were not pre-trained on dialogue data, we included DialogueLM LED (Zhong et al., 2022)— a variant of Longformer model with a 5,120-token input context length and was pre-trained on interview and radio conversation corpora—in our experiments.

The results measured against the human-labeled test set are presented in Table 13. While the fine-tuned Longformer models demonstrated performance comparable to GPT-4o across most dimensions, they showed a notable disparity in predicting the target speaker. This discrepancy may be attributed to the dynamic nature of classification labels—the number and identity of speakers change between episodes.

Generative or retrieval approaches are more effective for target speaker classification. Finally, we observed that pre-training the model with dialogue corpora did not noticeably impact performance.

## C Disagreement Cases Analysis

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

Table 14: 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 DEBATE dataset.

In this appendix, we highlight the complexity and difficulty of the task by curating several examples in Table 14. We analyze and discuss cases of disagreement, particularly within the DEBATE 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 5. 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.

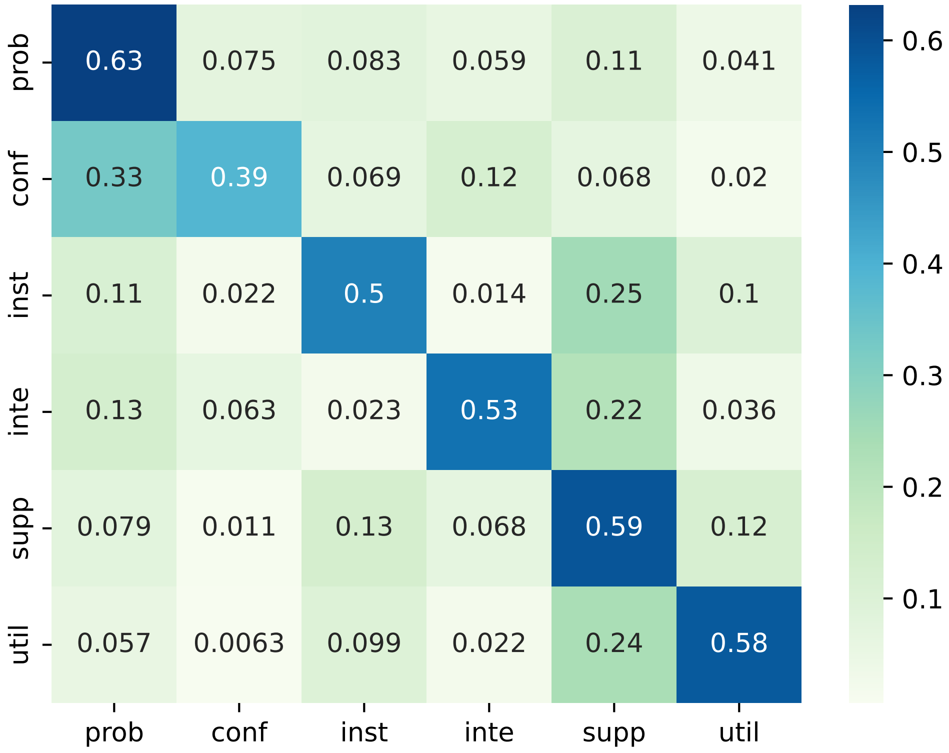


Figure 5: The normalized co-occurrence matrix of dialogue act human votes from the DEBATE subset.

## D Human Machine Annotation Comparative Analysis

| | | | DEBATE human | | | | |
|---|---|---|---|---|---|---|---|
| | prob | conf | inst | inte | supp | util | $p(m)$ |
| IM | **0.48*** | 0.09 | 0.05 | 0.25* | 0.11 | 0.02 | 0.37 |
| CM | 0.13 | 0.01 | 0.27 | 0.02 | **0.52*** | 0.04 | 0.53 |
| SM | 0.04 | 0.01 | 0.03 | 0.01 | 0.36* | **0.54** | 0.12 |
| Total | 0.24* | 0.04 | 0.15 | 0.09* | 0.34* | 0.14 | *150.77* |
| | | | DEBATE GPT-4o | | | | |
| IM | **0.40** | 0.22* | 0.04 | 0.11 | 0.22* | 0.01 | 0.39 |
| CM | 0.14 | 0.10* | **0.54*** | 0.02 | 0.10 | 0.11* | 0.66* |
| SM | 0.06 | 0.01 | 0.12* | 0.02 | 0.16 | **0.64** | 0.12 |
| $p(d)$ | 0.22 | 0.11* | 0.36* | 0.05 | 0.12 | 0.14 | *150.77* |
| | | | PANEL human | | | | |
| | prob | conf | inst | inte | supp | util | $p(m)$ |
| IM | **0.51*** | 0.02 | 0.02 | 0.02 | 0.42 | 0.01 | 0.60 |
| CM | 0.03 | 0.00 | 0.09 | 0.00 | **0.85*** | 0.03 | 0.28 |
| SM | 0.00 | 0.00 | 0.01 | 0.02 | 0.20 | **0.72** | 0.06 |
| $p(d)$ | 0.31 | 0.01 | 0.03 | 0.02 | 0.55* | 0.08 | *61.35* |
| | | | PANEL GPT-4o | | | | |
| IM | 0.41 | 0.04* | 0.01 | 0.03 | **0.50*** | 0.01 | 0.72* |
| CM | 0.08* | 0.02 | **0.41*** | 0.01 | 0.33 | 0.16* | 0.25 |
| SM | 0.05 | 0.00 | 0.02 | 0.01 | 0.27 | **0.64** | 0.16* |
| $p(d)$ | 0.30 | 0.03* | 0.10* | 0.02 | 0.41 | 0.13* | *61.35* |

Table 15: Conditional probabilities of dialogue acts (columns) given motives (rows), along with marginal probabilities of dialogue acts (right column) and motives (bottom row). All values are averaged across episodes from the **test** and **development** sets for the two scenarios—DEBATE (top) and PANEL (bottom)—and for the two annotation sources: human and GPT-4o. The most frequent dialogue act for each motive is highlighted in bold, with the second most frequent underlined. The italicized number in the corner indicates the average frequency of moderator sentences. An * denotes values that are statistically significantly greater than their annotation source counterparts (human vs. GPT-4o; t-test at $p <= 0.05$).

To validate the analytical findings from the automatic system, we conducted a comparative study between human annotations and machine-generated annotations (using GPT-4o) on the test and development datasets. Table 15 presents the conditional and marginal probabilities across the two settings (DEBATE vs. PANEL) and the two annotation approaches (human vs. GPT-4o). Overall, the results indicate that machine annotations generally align well with human annotations. Although some differences are statistically significant, their magnitudes are typically small ($\leq 0.1$).

One notable exception is the distinction between coordinative-motivated "instruction" and "supplement." In our error analysis, we found that this discrepancy arises from differences in interpreting the "immediacy" of the expected influence on subsequent turns. An "instruction" act is intended for moderator interventions that expect an immediate change in the target speaker's behavior (e.g., "Please stay on topic."). In contrast, when moderators provide information without expecting immediate action (e.g., "After the debate, we will proceed to voting."), human annotators tend to label it as a "coordinative-motivated supplement," as it provides context or rules without requiring an immediate response. Machine annotations, however, did not consistently capture this nuance and often mislabeled these explanations of rules as "instructions," overlooking the subtle difference in immediacy.

We acknowledge the need to refine these aspects of the annotation framework to improve accuracy. Nevertheless, the core patterns and characteristics identified by both human and machine annotations remain largely consistent, reinforcing the validity of our primary findings.

## E Classification Error Analysis

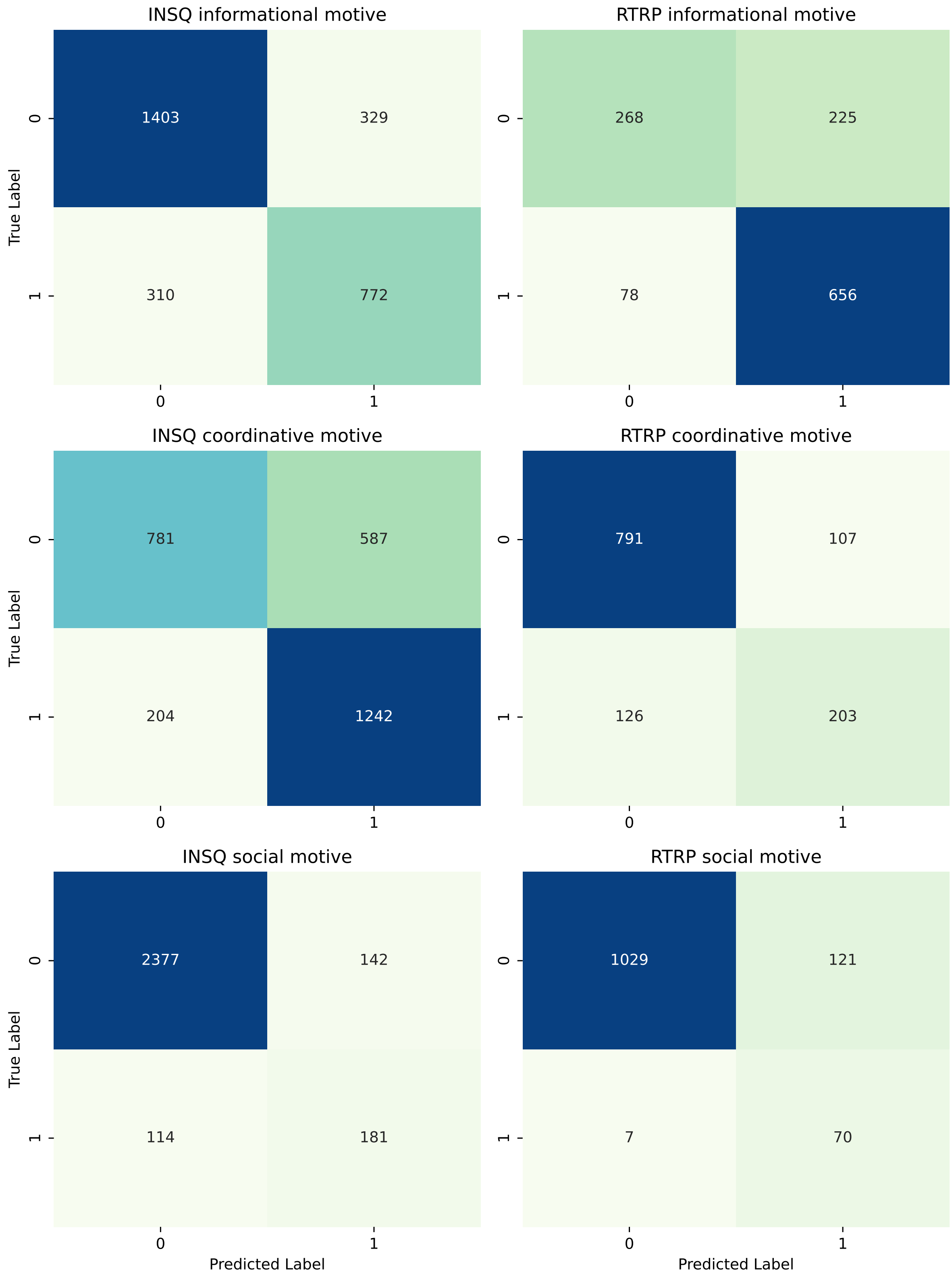


Figure 6: The confusion matrices for the three motives across the two subsets.

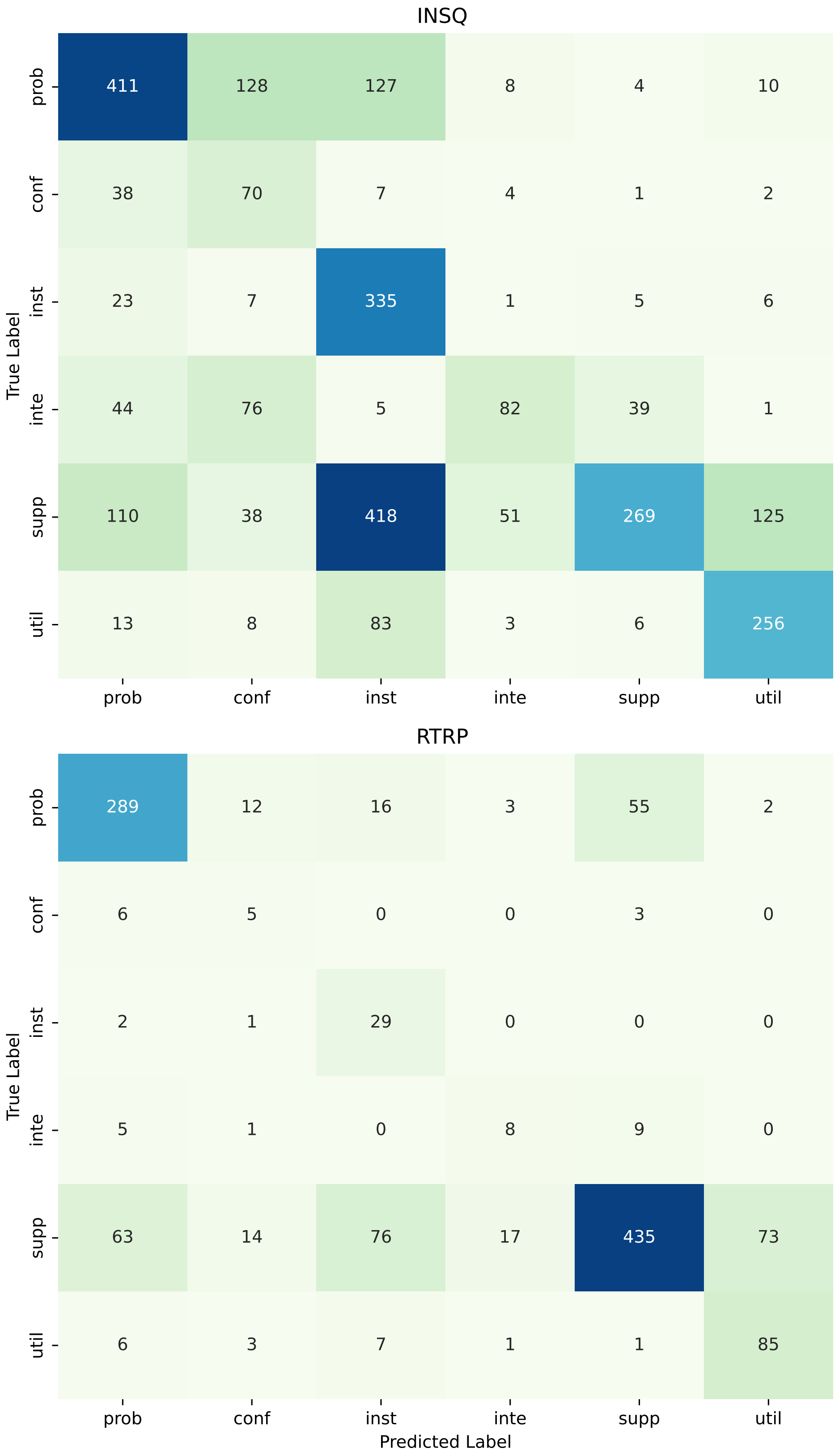


Figure 7: The confusion matrices for the three motives across the two subsets.

| Dimensions | Examples |
|---|---|
| Dialogue act | 1. Eli Pariser. (prob vs. conf, DEBATE).<br>2. Dr. David Satcher. (conf vs. prob, DEBATE).<br>3. I want to bring Matt back into this conversation. (prob vs. conf, DEBATE)<br>4. But wasn't your partner using the "that's what happened to me when I typed in Egypt"? (prob vs. inte, DEBATE)<br>5. Let's go to Frank Foer. (prob vs. inst, DEBATE)<br>6. There was a lot of questions that came up during Jena Six, saying, oh, marching is so 1965.(prob vs. supp, PANEL)<br>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, DEBATE)<br>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, DEBATE)<br>9. I have a– question in the second row. (supp vs. prob, DEBATE)<br>10. You work for the Washington Post and I couldn't even find the story online about that. (supp vs. prob, PANEL)<br>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, DEBATE)<br>12. Microphones will be brought forward if you raise your hand. (supp vs. inst, RTRO)<br>13. Yep (supp vs. util, DEBATE)<br>14. Alright (util vs. inst, DEBATE) |
| Motives | 15. So how would you relate that directly to the motion? (IM false positive, DEBATE)<br>16. Jacob Weisberg. (IM false negative, DEBATE)<br>17. What do you - Jasmyne, I'll start with you - unfold your, uncross your arms. (IM false negative, PANEL)<br>18. The team arguing against the motion, Franklin Foer and Scott Turow, they're saying, "It's all a trap. (CM false positive, DEBATE)<br>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, DEBATE)<br>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, PANEL)<br>21. Well, you're laughing because you think it's impossible or what is. . . (SM false positive, PANEL)<br>22. All right. (SM false negative, DEBATE) |
| 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, DEBATE)<br>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, DEBATE)<br>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, DEBATE)<br>26. Let me approach this from a couple of different angles. (all speakers vs. audience, PANEL) |

Table 16: 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 6 presents the confusion matrix for the three motives, comparing GPT-4o with the aggregated human annotations, while Figure 7 displays the confusion matrix for the six dialogue act labels. Table 16 provides examples of common errors across the three dimensions to support further qualitative analysis.

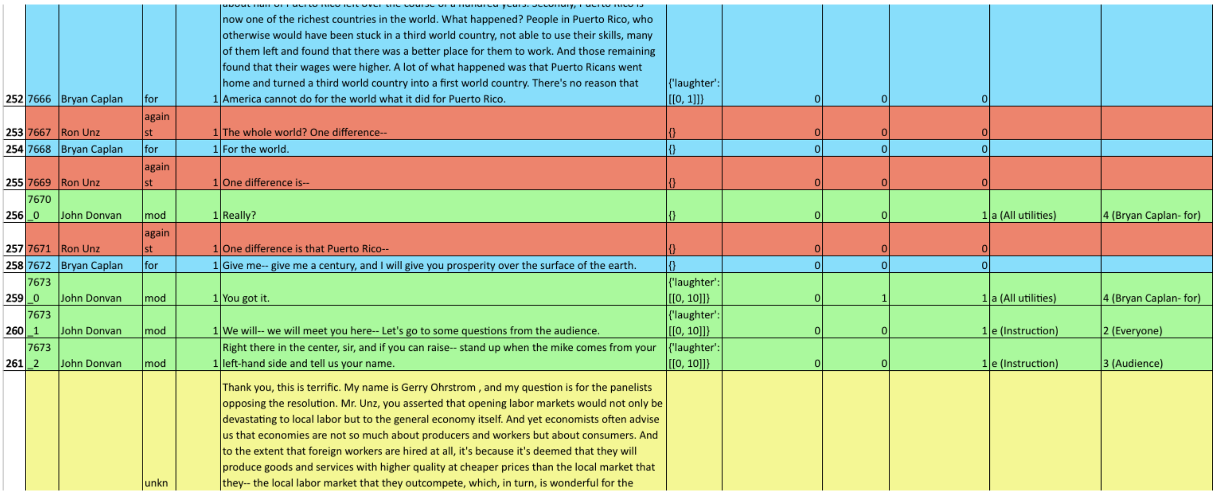
An analysis of the dialogue act confusion matrix in Figure 7, particularly within the DEBATE subset, reveals four primary sources of error. First, several probing sentences are frequently misclassified as confrontational or instructional. In Table 16, Examples 1 and 2 illustrate instances where the sentences 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 6, we identified two primary sources of error. In the DEBATE subset, the 'coordinative' motive exhibited the lowest performance, with most errors being false positives. For example, in Table 16, 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 PANEL 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.

## F Annotator instruction and material



| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 252 | 7666 | Bryan Caplan | for | 1 | about half of Puerto Rico left over the course of a hundred years. Secondly, Puerto Rico is 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 America cannot do for the world what it did for Puerto Rico. | {'laughter': [[0, 1]]} | 0 | 0 | 0 | | |
| 253 | 7667 | Ron Unz | again st | 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 | again st | 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 | again st | 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 8: 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.

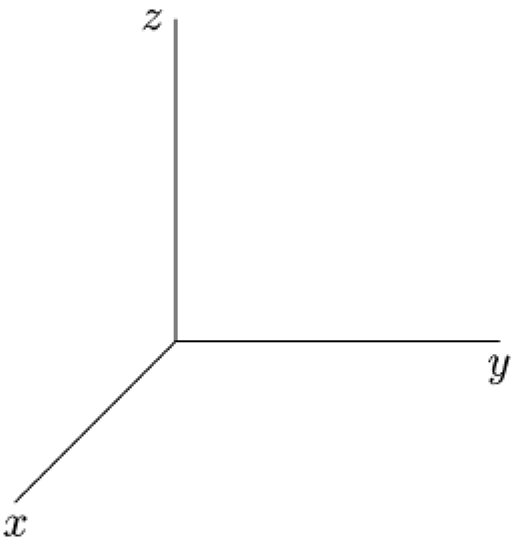
## ANNOTATION FACETS AND LABELS INTRODUCTION

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)

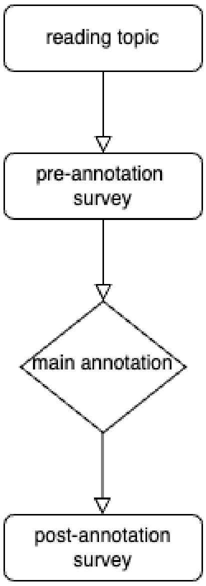
“Cough! Cough!” (Self)

“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. | | | |