Evaluation and Facilitation of Online Discussions in the LLM Era: A Survey

Katerina Korre[†], **Dimitris Tsirpas**[†‡], **Nikos Gkoumas**[†], **Emma Cabalé**[◇]
Danai Myrtzani[‡], **Theodoros Evgeniou**[♠]
Ion Androutsopoulos[‡†], **John Pavlopoulos**[‡†]

[†]Archimedes, Athena Research Center, Greece ({k.korre,n.goumas}@athenarc.gr)
[‡]Athens University of Economics and Business, Greece ({dim.tsirpas,dan.myrtzani,ion,annis}@aueb.gr)
[◇]École Normale Supérieure Paris-Saclay, France (emma.cabale@ens-paris-saclay.fr)
[♠]INSEAD, Technology and Business, France (theodoros.evgeniou@insead.edu)

Abstract

We present a survey of methods for assessing and enhancing the quality of online discussions, focusing on the potential of Large Language Models (LLMs). While online discourses aim, at least in theory, to foster mutual understanding, they often devolve into harmful exchanges, such as hate speech, threatening social cohesion and democratic values. Recent advancements in LLMs enable artificial facilitation agents to not only moderate content, but also actively improve the quality of interactions. Our survey synthesizes ideas from Natural Language Processing (NLP) and Social Sciences to provide (a) a new taxonomy on discussion quality evaluation, (b) an overview of intervention and facilitation strategies, (c) along with a new taxonomy of conversation facilitation datasets, (d) an LLM-oriented roadmap of good practices and future research directions, from technological and societal perspectives.

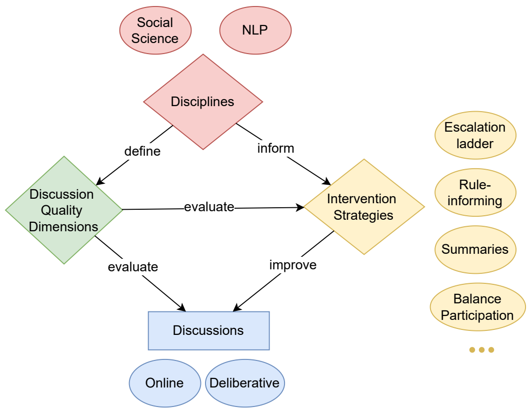

Figure 1: A conceptualization of this survey. We explore approaches from different disciplines, which recommend their own ways of evaluating and improving discussions.

1 Introduction

Discussions, especially of complex or controversial topics, are a cornerstone of collective decision-making (Burton et al., 2024). In contrast to initial hopes of promoting mutual understanding (Rheingold, 2000), online discussions (especially in social media) often degenerate into hate speech, personal attacks, promoting conspiracy theories or propaganda – to the extent that they can even be considered a threat to social cohesion and democracy (Tucker et al., 2018; Mathew et al., 2019).

Natural Language Processing (NLP) and Machine Learning (ML) can potentially help improve the quality of online discussions. For example, automatic classifiers (Bang et al., 2023; Molina and Sundar, 2022) are already being used to help or even replace human moderators, by flagging posts that violate the law or policies of online discussion fora (Saeidi et al., 2021).

Social Science provides theories and applications for the facilitation of a discussion, but in specific contexts, such as teaching (Mansour, 2024) or clinical discussions (Gelula, 1997), without much research devoted to online discussions. While prior NLP studies have explored LLM-facilitated discussions (Burton et al., 2024; Aher et al., 2023; Beck et al., 2024; Schroeder et al., 2024; Small et al., 2023; Cho et al., 2024), rarely does Social Science work examine how facilitation can be automated (Gimpel et al., 2024).

In this survey, we combine LLM-based methods, with ideas from Social Science (e.g., Deliberative Theory) when discussing how to evaluate online discussions, and when exploring intervention strategies. Figure 1 provides a high-level conceptualization of our work.

The main research question of this survey is *can LLMs be used effectively as facilitators in online discussions?* Focusing on threaded discussions (§2), we explore three key areas: (1) methods (potentially also LLM-based) for evaluating aspects of online discussions, (2) intervention strategies for facilitation, and (3) available data resources

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 24455–24474
November 4-9, 2025 ©2025 Association for Computational Linguistics

which can be used to analyze human facilitation and train LLM equivalents. Specifically, we survey discussion evaluation aspects and introduce a new taxonomy (§4). We map tasks suited for ML models, LLMs, and humans, aggregate multi-dimensional insights on facilitation strategies (§5), and outline future possibilities for LLMs (§6). Additionally, we aggregate and compare all major relevant datasets in literature, dividing them into categories per task (§7).

Our findings show that (a) many discussion evaluation dimensions coexist in the literature; (b) LLM advancements show significant promise in improving the quality and timeliness of facilitation methods; (c) while surveying the existing datasets, we notice a scarcity of datasets for studying facilitation. We posit that LLM-generated discussions, could become an asset to develop and test automatic facilitation strategies in diverse artificial discussions, before testing the strategies and the LLM-based facilitator agents in more costly experiments with human participants.

2 Terminology

Given the numerous aspects to consider regarding discussion quality and facilitation, we clarify the terminology we use. We highly recommend consulting the Terminology Section of Appendix C and, especially, Table 3, where we explain our findings with regard to the terms used in the literature.

Facilitation vs. Moderation The term 'moderation' is more commonly used in NLP (Argyle et al., 2023), typically referring to the flagging and/or removal of unwanted content ('content moderation'), while 'facilitation' is more prevalent in the Social Sciences, where it encompasses a broader scope, including active interventions (Vecchi et al., 2021; Kaner et al., 2007; Trenel, 2009). Given the limited attention to facilitation in NLP and the survey's grounding in Social Science, we distinguish between the terms, even though they are sometimes used interchangeably in the literature.

Ex-Post moderation This survey mainly focuses on 'Real-Time, Ex-Post-moderation', i.e., moderation happening just after the user has posted some content. This is different from pre-moderation approaches, such as nudging users before they post harmful content (Argyle et al., 2023), or delaying the posting of user content until a moderator has had the chance to check it.

Discussion, Deliberation, Dialogue, Debate The definitions of these terms often vary across literature (Russmann and Lane, 2016; Goñi, 2024). We focus on **discussions**, a general term for verbal/written exchanges (Russmann and Lane, 2016), and **deliberations**, a term for structured discussions focusing on opinion sharing (Degeling et al., 2015; Lo and McAvoy, 2023). This is in contrast to the (at least in theory) collaborative nature of **dialogues** (Rose-Redwood et al., 2018; Bawden, 2021; Goñi, 2024) and the competitive and organized nature of **debates** (Lo and McAvoy, 2023).

Tree-style discussions (or 'threads') are discussions which start from an Original Post (OP) with subsequent comments replying to either the OP or to other comments (Seering, 2020).

3 Comparison to Other Surveys

Only two studies have surveyed the field of NLP while also considering ideas from Social Science. However, they focus mainly on Argument Mining (AM). These are the studies of Wachsmuth et al. (2024) and Vecchi et al. (2021). Wachsmuth et al. (2024) focus primarily on discussion *evaluation* disregarding its relation to facilitation, which is one of the main goals of our survey. The survey of Vecchi et al. (2021) argues that advancing AM for social good requires a collaborative effort between AM and Social Science. They point out that traditional AM has prioritized the logical structure and soundness of arguments, while overlooking other important dimensions, such as civility, respectfulness, inclusiveness, originality, and the broader impacts of discussions—such as encouraging mutual understanding and problem-solving. Building on these notions, we incorporate ideas from Social Science into NLP-based approaches, discussing both discussion evaluation and facilitation, both with a focus on the potential of LLMs.

4 Discussion Quality Evaluation

Improving online discussions presupposes being able to define and measure *discussion quality*. While there have been attempts to provide frameworks for discussion quality evaluation (Kies, 2022; Gerber et al., 2018), none of them is directed towards facilitation. Crucially, most existing frameworks ultimately rely on human judgments as their reference point, yet human evaluation is expensive, slow, and shows low inter-rater agreement on dimensions that involve subjective interpretation, such as pragmatic cues (Smith et al., 2022; Yeh et al., 2021; Khalid and Lee, 2022). This evaluation bottleneck motivates a taxonomy of evaluation methods that is both comprehensive and amenable to scalable automatic measurement.

In this work, we draw from the works of Bächtiger et al. (2022, 2010); Steenbergen et al. (2003); Falk and Lapesa (2023) and Kies (2022) to define a new social-science-informed taxonomy for discussion quality dimensions. While we present a structured taxonomy, it is important to note that the categories are not mutually exclusive. Rather, elements within the taxonomy may coexist within evaluation dimensions, complement one another, or serve as explanatory mechanisms for other dimensions. An example of the dimension interaction can be found in Table 5 in the Appendix F. The grouped dimensions along with the NLP approaches are shown in the Appendix in Table 4.

4.1 Structure and Logic

Argument Structure and Analysis Argument Quality (AQ) is a multidimensional concept assessed through logical, rhetorical, and dialectical dimensions (Wachsmuth et al., 2017). The logical dimension focuses on the coherence and structure of the argument. The rhetorical dimension assesses persuasiveness, focusing on the argument's style and emotional appeal. The dialectical dimension assesses the constructiveness of the argument. Empirically, threads with well-formed claim-evidence chains exhibit higher coherence and lower odds of devolving into ad-hominem attacks, making AQ scores, as a discussion quality dimension, an *early-warning* indicator of derailment (Chang and Danescu-Niculescu-Mizil, 2019). All the above dimensions of automatic argument-structure analysis can be used by a facilitator to keep the discussion fact-centered, inclusive, and on track (Falk et al., 2021; Falk and Lapesa, 2023).

Coherence and Flow Coherence, as described above, evaluates logical consistency, while flow assesses smooth progression in discussions (Li et al., 2021). Both are essential tools for facilitators in their effort to redirect off-topic comments and guide transitions between topics during a discussion (Lambert et al., 2024; Park et al., 2012; Falk et al., 2024). A sudden drop in how well responses match the topic or question often comes before personal attacks or off-topic turns (Chang and Danescu-Niculescu-Mizil, 2019; Zhang et al., 2018), making coherence and flow indicators of argument structure and a valuable early signal for facilitators.

Turn-taking How speakers alternate, the frequency of their turns, and the participants they address can serve as a diagnostic of conversational health. Balanced exchanges enhance coherence (Cervone and Riccardi, 2020), predict constructiveness (§4.3) (Niculae and Danescu-Niculescu-Mizil, 2016), and provide facilitators with actionable cues (Schroeder et al., 2024). To gauge speaking time, turn count, and word usage, researchers have applied metrics such as entropy (Niculae and Danescu-Niculescu-Mizil, 2016) and Gini coefficients (Schroeder et al., 2024).

Linguistic Markers Linguistic markers have been used to help model content and expression in online discussions (Wilson et al., 1984). Early methods used lexicons for sentiment, toxicity, politeness (§4.2 and 4.3) and collaboration evaluation (Lawrence et al., 2017; Avalle et al., 2024). For example, spikes in hedges (e.g., 'maybe', 'I guess') invite clarification requests by facilitators, while bursts of second-person pronouns, similarly to turn-taking, often foreshadow personal attacks and can prompt a civility nudge (Niculae and Danescu-Niculescu-Mizil, 2016).

Speech and Dialogue Acts Rooted in Speech Act Theory (Austin, 1975; Searle, 1969), dialogue acts have been employed to assess deliberative quality and analyze facilitation strategies (Fournier-Tombs and MacKenzie, 2021; Chen et al., 2024). They characterize dialogue turns (e.g., interruption) to analyze interaction dynamics (Ferschke et al., 2012; Stolcke et al., 2000; Zhang et al., 2017; Al-Khatib et al., 2018). Positive (e.g., causal reasoning) or negative (e.g., disrespect) dialogue acts can be scored to reflect discussion quality with low scores potentially indicating a need for intervention (Ziems et al., 2024; Cimino et al., 2024; Martinenghi et al., 2024; Schroeder et al., 2024).

Pragmatic Comprehension Pragmatic comprehension—how context shapes meaning—is crucial to facilitation, as intended meanings often diverge from literal expressions (i.e., implicature). Humans resolve such ambiguity using social and commonsense knowledge. Grice's maxims (Grice, 1975), a central pragmatic concept, can help explain this process by outlining the conversational principles people rely on to infer meaning, while they have already been used to assess discussion quality (Jwalapuram, 2017; Langevin et al., 2021; Ngai et al., 2021; Nam et al., 2023).

4.2 Social Dynamics

Politeness Politeness serves as a cornerstone of prosocial behavior, an attribute that facilitators desire to foster in online discussion forums (Lambert et al., 2024). In the context of facilitation, it has mainly been studied in relation to conversational derailment (§7) (Zhang et al., 2018) and constructiveness (§4.3) (De Kock and Vlachos, 2021; Zhou et al., 2024).

Power and Status Power and status influence conversational dynamics, affecting language use and turn-taking (§4.1). Higher status speakers can control the flow of discussions and foster social inequalities. Interestingly, low-status individuals tend to mimic the linguistic styles of high-status speakers more than the opposite (Danescu-Niculescu-Mizil et al., 2012), and this can be used as a signal that there is high/low-status imparity in a discussion. Facilitators may intervene, then, to ensure that the right to speak is evenly distributed among participants, preventing projection of social biases and stereotypes.

Disagreement Disagreements, when constructive, improve discussions by fostering deeper understanding (Friess, 2018; De Kock and Vlachos, 2021). Assessing disagreement, however, is complex. The hierarchy of Graham (2008) considers disagreement tactics ranging from name calling to refuting the central point. Along with other work on dispute tactics (Walker et al., 2012; Benesch et al., 2016; De Kock et al., 2022), it can be used to examine types of disagreements in a discussion.

4.3 Emotion and Behavior

Empathy Empathy is the ability to understand other perspectives and emotions and respond correspondingly (Lipman, 2003; Xu and Jiang, 2024). Facilitators desire to foster empathy in online discussions, since it encourages prosocial behavior and boosts engagement (Xu and Jiang, 2024; Concannon and Tomalin, 2024; Lambert et al., 2024). To do so, they encourage users to share personal stories and experiences (Schroeder et al., 2024). Various coding schemes (Macagno et al., 2022), psychological indicators (e.g., the emotion-laden words of Furniturewala and Jaidka, 2024), and dimensions (e.g., perceived engagement such as in Xu and Jiang, 2024) have been used to detect both expressed and perceived empathetic traits.

Toxicity Toxicity in online discussions refers to harmful or disrespectful language that hinders productive discourse and can derail meaningful discussions (Avalle et al., 2024). Facilitation is key to maintaining healthy communication, requiring both early detection of toxicity and (in the case of more active facilitation) proactive de-escalation strategies, such as conversation redirection or positive engagement (§5). In the case of conventional moderation that only aims to flag or remove toxic content, debate persists over what content warrants removal (Warner et al., 2025; Habibi et al., 2024; Pradel et al., 2024).

Sentiment Sentiment analysis helps identify whether discussions are positive, negative, or neutral. In the context of facilitation, sentiment analysis gauges the tone of discussions, which influences the quality of interactions (De Kock and Vlachos, 2021). Positive sentiment contributions in online discussion forums usually signal prosocial behavior and hence are highly encouraged by facilitators (Lambert et al., 2024), while negative sentiments among discussants contribute to conversation toxicity (Avalle et al., 2024).

Controversy Controversy arises from divergent viewpoints, leading to polarized exchanges that can escalate to toxicity and derail online discussions (Avalle et al., 2024). Controversial comments have been shown to contribute to a decline in positive emotions and a sustained rise in anger (Hessel and Lee, 2019; Chen et al., 2025). The spread of political leanings among discussants and sentiment distribution analysis are common approaches to measure controversy (Avalle et al., 2024).

Constructiveness Constructiveness fosters meaningful dialogue, especially in online discussions, by promoting resolution and cooperation (Shahid et al., 2024). It is often signalled by linguistic markers (§4.1) (De Kock et al., 2022; Falk et al., 2024). A facilitator can exploit a constructiveness score; threads trending upward are worth highlighting or summarizing, whereas a downward drift may trigger facilitation tactics such as slower, structured turn-taking or clarification prompts (De Kock and Vlachos, 2021).

4.4 Engagement and Impact

Engagement Engagement is desirable in online discussion platforms as it combines interest and participation (Lambert et al., 2024; Park et al., 2012). It is proxied by measures like reciprocity (Graham and Witschge, 2003; Stromer-Galley, 2007; Zhang et al., 2018), number of comments posted by each user (Avalle et al., 2024), discussion length (Adomavicius, 2021; Avalle et al., 2024), while Ferron et al. (2023) define subdimensions such as response diversity, interestingness, and specificity.

Persuasion Empirical literature has primarily examined factors influencing persuasion that align with other categories in our taxonomy, such as linguistic markers (§4.1) and turn-taking (§4.2) (Tan et al., 2016). Considering this connection, persuasion is not only an indicator of argument quality, but may also serve as a proxy for identifying additional markers signaling whether facilitator intervention is needed.

Diversity and Informativeness Diversity in online discussions refers to the presence of varied perspectives, backgrounds, and experiences, which can enrich conversations by fostering constructive exchanges (Irani et al., 2024; Zhang et al., 2024). To prevent echo chambers and promote inclusivity, facilitators can use diversity measures to encourage opinion diversity (Anastasiou et al., 2023), encouraging users to explore a broad range of perspectives on a given issue (Kim et al., 2021). Informativeness refers to the relevance and value of information shared in a discussion and is considered a building stock of prosociality, an attribute that facilitation trys to foster in online discussion platforms (Lambert et al., 2024).

4.5 LLM Approaches to Discussion Quality

LLMs can significantly aid in evaluating discussion quality, performing on par with humans in annotating argument structure (Mirzakhmedova et al., 2024; Rescala et al., 2024), excelling in comparative argument evaluation (Wang et al., 2023), AM, and synthesis (Chen et al., 2024; Irani et al., 2024; Anastasiou and De Liddo, 2024). They are increasingly used for coherence evaluation at the comment or whole discussion level (Zhang et al., 2024), often using proprietary models (e.g., GPT-4), while fine-tuned open-source models also show promise (Mendonca et al., 2024; Zhang et al., 2023). LLMs are not preferred for turn-taking or linguistic markers. Research on the former focuses on visual dashboards (such as VisArgue or TurnViz) that reveal dominance shifts at a glance (El-Assady et al., 2017; Hoque and Carenini, 2016), while distinguishing linguistic markers is often approached through older methodologies such as LSTMs (Sak et al., 2014) or dictionaries, as mentioned in §4.1.

LLMs can also serve as dialogue and speech act annotators (Ziems et al., 2024; Cimino et al., 2024; Martinenghi et al., 2024; Schroeder et al., 2024). For example, Yu et al. (2024) show that GPT-4 reached almost human accuracy on the task of annotating the speech act of apologizing. However, we acknowledge that the difficulty of automatic speech act annotation might depend on the task and more research on that is encouraged.

Remaining on the frontier of pragmatics, research shows that LLM-fine-tuning enhances implicature comprehension (Ruis et al., 2023), with GPT-4 achieving human-level performance through chain-of-thought prompting. While LLMs perform well in some pragmatics tasks, such as in the Pragmatic Understanding Benchmark (PUB) (Sravanthi et al., 2024), they struggle with social norm-based understanding (e.g., humor, irony) (Hu et al., 2023; Sravanthi et al., 2024). This is also true for annotating politeness, power, disagreement, and toxicity (Zhou et al., 2024; Ziems et al., 2024).

LLMs perform well in identifying power differentials in discussions (Ziems et al., 2024), and can detect these imbalances in real time, enabling facilitators to invite quieter voices and limit dominant turns. Additionally, LLMs have been successfully employed as dispute tactics annotators, highlighting instances of hostile interactions that may require moderator intervention (Zhou et al., 2024). However, they show limited accuracy in sentiment and engagement detection (Hu et al., 2023; Sravanthi et al., 2024; Furniturewala and Jaidka, 2024; Xu and Jiang, 2024). Empathy detection also remains challenging for LLMs, with evaluations showing inconsistent performance across conversational tasks (Furniturewala and Jaidka, 2024; Xu and Jiang, 2024; Ziems et al., 2024). While LLMs show promise in measuring controversy and persuasion, performance drops at the discussion level, particularly when assessing diversity, informativeness, and broader aspects of sociopragmatic understanding (Ziems et al., 2024; Avalle et al., 2024; Lawrence and Reed, 2020).

5 Intervention Strategies

5.1 When to Intervene

Picking the right moment to intervene is a crucial part of effective facilitation strategies. If a facilitator does not intervene when they should have, there is a risk of significant escalation, while intervening when unnecessary can increase toxicity (Schaffner et al., 2024; Trujillo and Cresci, 2022; Schluger et al., 2022; Cresci et al., 2022). Even 'softer' interventions such as information and opinion sharing can prove detrimental to discussions when performed excessively (Gao et al., 2025). It is imperative then for a facilitator to be able to recognize subtle cues that hint towards escalation (also considering the evaluation dimensions discussed in §4), in order to defuse the situation, something that even experienced human facilitators are not confident to reliably do (Schluger et al., 2022).

The NLP task of 'Conversational Forecasting' may contribute towards this direction. Given a conversation up to a point, a model attempts to predict if an event will occur in the future in that conversation. In our case, this is where a facilitator would intervene (Schluger et al., 2022). Traditional ML models can perform well on this task, although their performance varies (Falk et al., 2021; Park et al., 2012; Falk et al., 2024; Schluger et al., 2022).

5.2 How to Intervene

There is currently no agreed-upon taxonomy for facilitator interventions. Lim et al. (2011) propose a taxonomy that focuses on discussion facilitation, excluding, however, disciplinary or administrative actions, which are common in online discussions. Park et al. (2012) propose another taxonomy consisting of seven moderator functions, ranging from policing the discussion to solving technical issues. Their taxonomy, however, is not easily generalizable to domains other than website facilitation (Chen et al., 2025). These functions roughly correlate with the volunteer moderator roles, as described by Seering (2020). More practical approaches can be found in facilitator manuals (eRulemaking Initiative, 2017; MIT Center for Constructive Communication, 2024) and books (White et al., 2024). Chen et al. (2025) bridge the questions of *when* and *how* to facilitate by proposing a taxonomy that analyzes both individually, which was improved by Gao et al. (2025).

Several works have examined facilitation in education. Sjølie et al. (2021) conducted a mixed-methods study on a meta-facilitative approach, where students and teachers explicitly discussed their collaboration and which led to significant learning improvements. In the context of virtual facilitation, Verkuyl et al. (2024) showed that successful integration of virtual simulations in higher education depends not just on access, but on facilitators who align simulations with course objectives, respond to learners' needs, and evaluate the experience. Both studies suggest that facilitation requires socially informed practices, even as automation promises workload reduction.

With reference to NLP approaches in facilitation in education, Lugini et al. (2020) designed Discussion Tracker, a classroom analytics tool that applies algorithms to identify argumentation moves (claim, evidence, explanation) and evaluate levels of specificity, as well as recognize patterns of collaboration. Deployment in class showed that teachers considered the analytics valuable, and that the system's classifiers achieved moderate to substantial agreement with human judgments. The aforementioned work of Gao et al. (2025) presented an approach that combines automatic English as a Second Language (ESL) dialogue assessment with a framework of moderation strategies. The authors showed that moderators improve topic flow and conversation management, with active acknowledgment and encouragement proving most effective, but excessive input can hinder discussion.

Facilitators often have to decide what form of coercive measure to take to make sure the conversation remains healthy, without having to intervene repeatedly. Human interventions typically use an unofficial 'escalation ladder' (Figure 1), where the facilitator will progressively move from milder facilitation tactics to threatening, and finally enacting disciplinary action (Seering, 2020). 'Conversational moderation' (Cho et al., 2024), where a facilitator first converses with the offender, has proven effective and is actively encouraged in some facilitator guidelines (The Commons, 2025). This is probably why disciplinary action is typically not the first choice of a facilitator (Schluger et al., 2022) and why it should reasonably be used as a last resort.

Softer kinds of interventions that facilitators frequently use first include: setting and informing users about rules (Schluger et al., 2022; Seering, 2020), welcoming new users (Schluger et al., 2022), summarizing key points (Small et al., 2023; Falk et al., 2024), balancing participation (Kim et al., 2021; Fishkin et al., 2018), and aiding users improve their points (Tsai et al., 2024; Falk et al., 2024). Facilitators are also instrumental in beginning and ending discussions (Small et al., 2023; Gao et al., 2025), as well as generally encouraging participants (Gao et al., 2025). It is worth noting that facilitation guides may explicitly forbid facilitators from intervening in certain ways, such as sharing their opinions or providing new information (MIT Center for Constructive Communication, 2024).

5.3 Personalized Interventions

Intervention strategies should not be applied en masse, without considering the characteristics of each individual. Traditionally, massive application of disciplinary action (or threatening) has led to adverse effects community- and platform-wide (Trujillo and Cresci, 2022; Falk et al., 2021) and to the creation of echo-chambers (Cho et al., 2024). There are also calls for research to move away from one-size-fits-all approaches and instead move towards personalized interventions (Cresci et al., 2022). Human facilitators are often able to personalize interventions per individual (Schluger et al., 2022), and we hypothesize that LLMs can also do so to some extent.

6 Towards LLM-based facilitation

Until recently, ML models used as facilitation agents were confined to either performing menial tasks, such as pasting automated messages (Seering, 2020; Schluger et al., 2022), suggesting facilitation actions (e.g., rejecting posts), possibly via human-in-the-loop frameworks (Fishkin et al., 2018; Gelauff et al., 2023), identifying possibly escalatory comments (Schluger et al., 2022), or employing pre-programmed facilitative tactics, as in the work of Kim et al. (2021), where the model produces automated messages encouraging participation. However, older ML-based and rule-based facilitation are not effective enough to meet the high demands of most platforms (Seering, 2020; Schaffner et al., 2024).

Advances in LLMs enable the development of *facilitation agents* that engage more actively in discussions. These agents can warn users about policy violations (Kumar et al., 2024), suggest rephrasings to improve tone or persuasiveness (Bose et al., 2023), monitor turn-taking (Schroeder et al., 2024), and summarize or visualize key discussion points

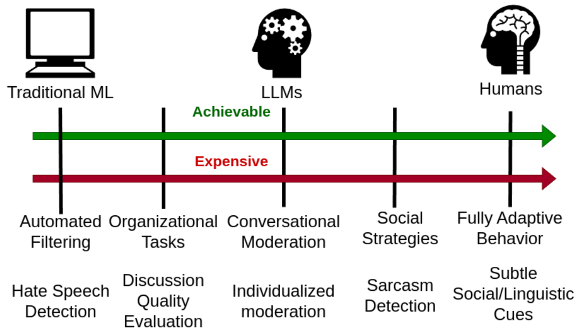

Figure 2: Capabilities of simpler ML, LLM, and human facilitation. Task complexity and cost increase from left to right. Intermediate tasks are handled suboptimally by the preceding method.

(Small et al., 2023). They can also assist in drafting group statements that reflect diverse viewpoints (Tessler et al., 2024). A brief, non-exhaustive summary of the capabilities of simpler ML models, LLMs, and humans can be found in Figure 2.

6.1 Administrating the Discussion

LLMs are able to tackle a variety of 'administrative' facilitation tasks that help structure discussions. For example, facilitators often summarize the views of the participants, seek confirmation of understanding, and share perspectives. This iterative summarization is a task LLMs may handle effectively (Small et al., 2023; Burton et al., 2024). However, Feng and Qin (2022) point out some challenges such as discussions with multiple participants, topic drifts, multiple co-references, diverse interactive signals, and diverse domain terminologies. Still, according to Jin et al. (2024), LLMs bring significant advantages over conventional ML methods, "notably in the quality and flexibility of the generated texts and the prompting paradigm to alleviate the cost of training deep models".

In some deliberative contexts, facilitators are also encouraged to begin a discussion with their own opinion (Small et al., 2023), although others disagree (MIT Center for Constructive Communication, 2024). This is a task LLMs can also handle, albeit less convincingly than Information Retrieval (IR) approaches (Karadzhov et al., 2021).

Finally, LLMs can help marginalized groups in discussions by offering translations of the discussions in their native languages, and by helping them phrase their opinions with proper grammar and syntax (Tsai et al., 2024; Burton et al., 2024). This can directly improve discussions by increasing their diversity (Section 4.4).

6.2 Evolving Traditional Automation Models

LLMs have been proven to be adept at NLP tasks such as the detection of hate speech (Shi et al., 2024), toxicity (Kang and Qian, 2024; Wang and Chang, 2022), and misinformation (Kang and Qian, 2024; Wang and Chang, 2022). These abilities make LLMs usable as drop-in replacements for traditional ML models for these tasks, suggesting that conversational LLM facilitation agents may be able to identify, and dynamically adapt to such phenomena properly. We note however that LLMs are much more expensive and less scalable than their simpler ML counterparts. Furthermore, LLM annotation has its own challenges: LLM survey responses (Jansen et al., 2023; Bisbee et al., 2024; Neumann et al., 2025) and annotations (Gligori'c et al., 2024) are generally unreliable and surface-level. Non-deterministic behavior is also common in LLMs (Atil et al., 2025), but also particularly in closed-source models (Bisbee et al., 2024) on which a lot of research on LLM annotation hinges.

6.3 Fully Automated LLM-based Facilitation

There are indications that LLMs can be used as facilitators to the fullest capacity of the role. LLMs are able to predict optimal facilitation tactics (Schroeder et al., 2024), like traditional ML models (Al-Khatib et al., 2018). Furthermore, they have proven capable of developing and executing social strategies in other tasks, e.g., negotation games, LLM interactions (Abdelnabi et al., 2024; Cheng et al., 2024a; Martinenghi et al., 2024). Given that relatively simple ML chatbots, which do not leverage generative text capabilities, have been reported to improve discussions (Kim et al., 2021), many expect LLM-based facilitation to be a promising solution to the well-known bottleneck of human facilitation (Small et al., 2023; Seering, 2020; Burton et al., 2024; Schroeder et al., 2024). Notably, Cho et al. (2024) successfully use LLM facilitators with prompts based on Cognitive Behavioral Therapy to moderate a live discussion with human participants. Their work shows that LLM facilitators can adapt their instructions to users, although they cannot by themselves affect the discussion with regard to cooperation and mutual respect between the participants.

Nevertheless, LLMs have inherent limitations that make them worse than humans in most social tasks (Figure 2; Rossi et al. (2024)). While human facilitators are encouraged to be neutral (White et al., 2024; eRulemaking Initiative, 2017), numerous studies point to biases in sociodemographic, statistical, and political terms in LLMs (Anthis et al., 2025; Hewitt et al., 2024; Rossi et al., 2024), which can be exacerbated during the course of a discussion (Taubenfeld et al., 2024).

7 Facilitation Datasets

In this section, we provide an overview of the most prominent datasets for online facilitation, considering their sizes and their relevance to core facilitation tasks. These datasets can be used for analyzing the behavior of human facilitators and the reactions of the participants, investigating the existing taxonomies (e.g., ones presented in §5) or as training data for human and LLM facilitators.

Due to the low number of such datasets in literature (Chen et al., 2025), the entries presented in this section straddle various domains adjacent to online facilitation. Hence, we propose the following new taxonomy of facilitation datasets: *Conversation Derailment* datasets, where the task is to predict when a conversation escalates, therefore requiring facilitator intervention; and *Facilitator Interventions* datasets, which include comments by facilitators in active discussions, sometimes annotated with the tactics employed. Some datasets contain information that can be used in multiple tasks.[1] An overview of the surveyed datasets and their categories in our taxonomy can be found in Table 1.

8 LLM Discussion Facilitation Roadmap

Evaluation LLMs can serve as automated discussion quality annotators (§4). Are these annotators infallible? Not yet. Certain dimensions, especially those that are highly subjective (e.g., pragmatic understanding), remain challenging for LLMs to annotate accurately. But we must take into account that even human annotations tend to be polarized for such subjective quality dimensions (Argyle et al., 2023), largely due to sociodemographic background effects and personal biases (Beck et al., 2024; Sap et al., 2020).

On the other hand, prompted LLMs offer a more scalable and cost-effective alternative for annotating discussion quality compared to human annotation and traditional (or self-) supervised training

[1]Despite its designation, the 'WikiDisputes' dataset does include information about facilitators. We consider it solely a 'Conversation Derailment' dataset because facilitator interventions only constitute 0.03% of its comments.

Name	Task		Size	Content
WikiDisputes (De Kock and Vlachos, 2021)	Conversation Derailment		7,425 D	Includes annotations for several 'dispute tactics.'
Wiki-Tactics (De Kock et al., 2022)	Conversation Derailment	Facilitator Interventions	213 D	Based on Wikipedia Disputes, includes moderation action metadata such as comment edits and deletions.
WikiConv (Hua et al., 2018)	Facilitator Interventions		91,000,000 D	Includes moderation meta-data such as comment edits and deletions.
Conversations Gone Awry (Zhang et al., 2018)	Conversation Derailment		4,188 D	Predicts derailment by analyzing rhetorical tactics, human-annotated.
Chang and Danescu-Niculescu-Mizil (2019) (1)	Conversation Derailment		4,188 D	Extends the 'Conversations Gone Awry' dataset.
Chang and Danescu-Niculescu-Mizil (2019) (2)	Conversation Derailment		6,842 D	Based on the r/ChangeMyView subreddit.
Park et al. (2012)	Conversation Derailment	Facilitator Interventions	1,678 C	Comprised of 4 datasets. Includes 19 intervention types belonging to 7 moderator roles.
RegulationRoom (Falk et al., 2021)	Conversation Derailment		3,000 C	Extends the dataset of Park et al. (2012).
UMOD (Falk et al., 2024)	Facilitator Interventions		2,000 C	Based on the r/ChangeMyView subreddit, annotated for facilitation tactics and AQ.
Fora (Schroeder et al., 2024)	Facilitator Interventions		262 D	Original dataset revolving around experience-sharing, annotated for facilitation tactics.
WHoW (Chen et al., 2025)	Facilitator Interventions		21,151 C	Dataset derived from TV debates and radio panels, annotated for facilitation tactics.
L2Moderator (Gao et al., 2025)	Facilitator Interventions		17 D (16.5 hours of transcripts)	Facilitated online discussions for ESL speakers.

Table 1: Overview of reviewed datasets. Unnamed datasets are referred to by the names of the authors only. The size reflects the number of annotated conversations, disregarding unlabeled data. **D** indicates the number of discussions. **C** indicates the number of individual comments or dialogue turns.

on large annotated datasets. Using LLMs for annotation, however, requires careful model selection considering whether models are open or closed source, model size, model alignment, as well as prompt selection, and (if applicable) fine-tuning requirements. These choices should be tailored to the specific quality dimension being evaluated.

Facilitation Intervention types should be adapted to the different legal frameworks, rules, and social norms of each community/platform. While there are exhaustive surveys on intervention types and policies, such as that of Schaffner et al. (2024), there is yet no methodology to train human or artificial facilitators according to these factors. We posit that experiments using exclusively LLM user/facilitator-agents are necessary to sustainably test new facilitation strategies and interventions per community and platform, as in other NLP tasks that involve LLM-generated conversation (Ulmer et al., 2024; Cheng et al., 2024b; Park et al., 2022, 2023), before testing the resulting facilitators in costly experiments with human participants. Finally, the datasets presented in Table 1 can be used to train and assess LLM facilitators in the future, as well as to generate additional data—similar to the existing ones, but with controlled modifications—to stress-test various facilitators in particular settings (e.g., predicting or recovering from a conversation derailment).

9 Conclusions

This survey examined online discussion evaluation and facilitation by bridging insights from Social Science and NLP, with a focus on the growing role of LLMs. We introduced a new discussion evaluation taxonomy, with categories that should remain flexible depending on the evaluation task and the characteristics of the discussion. In terms of intervention strategies, both human- and machine-driven advancements show significant promise in improving the quality of interventions, helping online discussions remain constructive, and resistant to derailment. Most facilitation datasets still originate from human online conversations, with research yet to fully explore the capabilities of LLMs. Taking the above into account, we believe that now is the time to embrace LLMs for facilitation to foster healthier and more constructive conversations.

10 Limitations

This survey is not without its limitations. While we have attempted to present a comprehensive overview of facilitation methods, certain techniques, such as summarization, could be explored in greater depth. Since summarization is a vast subfield of NLP, it was only briefly mentioned.

Moreover, it is important to highlight that most research on facilitation has been conducted solely in English-speaking online spaces. The inherent limitations of LLMs in handling other languages and cultural contexts must be considered. As a result, these findings may not be easily applicable to other regions of the world.

Finally, the majority of real-world online discussions and deliberations happen in the context of communities, where group dynamics (social behaviors, power structures, norms, and interactions) apply. Thus, a fuller review of facilitation would have to account for the internal dynamics of such communities, as well as the wider role of the facilitator as a figure that not only helps in the conversation but has a social status in the group as well.

11 Ethical Considerations

Although AI, and LLMs in particular, can be effectively used as discussion facilitators, offering dynamic, responsive discussion support, their deployment must meet strict transparency, safety, and accountability standards, especially for high-risk applications, as stated in the EU AI Act.[2] For example, a person or minority group may have been unfairly disadvantaged in an AI-enhanced deliberation. It is also necessary for the users to be aware that they are interacting with AI facilitators. Ideally, the consent of the users should be sought before using any sort of AI-enhanced discussion platform.

Even if LLMs facilitators eventually achieve a high level of autonomy, it is advisable to maintain human oversight. Keeping a human-in-the-loop approach ensures greater transparency and enables effective error prevention, detection, and correction.

Acknowledgments

This work has been partially supported by project MIS 5154714 of the National Recovery and Resilience Plan Greece 2.0 funded by the European Union under the NextGenerationEU Program.

[2] https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai

References

S. Abdelnabi, A. Gomaa, S. Sivaprasad, L. Schönherr, and M. Fritz. 2024. Cooperation, competition, and maliciousness: LLM-stakeholders interactive negotiation. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, pages 96–106, Vancouver, Canada.

S. Adomavicius. 2021. Putting the social in social media: How human connection triggers engagement. In *Proceedings of the New York State Communication Association*, volume 2017.

G. Aher, R.I. Arriaga., and A.T. Kalai. 2023. Using large language models to simulate multiple humans and replicate human subject studies. In *Proceedings of the 40th International Conference on Machine Learning*, pages 337 – 371, Hawaii, USA.

K. Al-Khatib, H. Wachsmuth, K. Lang, J. Herpel, M. Hagen, and B. Stein. 2018. Modeling deliberative argumentation strategies on Wikipedia. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2545–2555, Melbourne, Australia.

L. Anastasiou and A. De Liddo. 2024. A hybrid human-AI approach for argument map creation from transcripts. In *Proceedings of the First Workshop on Language-driven Deliberation Technology (DELITE)@ LREC-COLING 2024*, pages 45–51, Turin, Italy.

L. Anastasiou, A. De Moor, B. Brayshay, and A. De Liddo. 2023. A tale of struggles: an evaluation framework for transitioning from individually usable to community-useful online deliberation tools. In *Proceedings of the 11th International Conference on Communities and Technologies*, C&T '23, page 144–155, New York, NY, USA. Association for Computing Machinery.

J. R. Anthis, R. L., S. M. Richardson, A. C. Kozlowski, B. Koch, J. Evans, E. Brynjolfsson, and M. Bernstein. 2025. Llm social simulations are a promising research method. *Preprint*, arXiv:2504.02234.

L. P. Argyle, C. A. Bail, E. C. Busby, J. R. Gubler, T. Howe, C. Rytting, T. Sorensen, and D. Wingate. 2023. Leveraging AI for democratic discourse: Chat interventions can improve online political conversations at scale. *Proceedings of the National Academy of Sciences*, 120(41):1–8.

C. S. C. Asterhan and B. B. Schwarz. 2010. Online moderation of synchronous e-argumentation. *International Journal of Computer-Supported Collaborative Learning*, 5:259–282.

B. Atil, S. Aykent, A. Chittams, L. Fu, R. J. Passonneau, E. Radcliffe, G. R. Rajagopal, A. Sloan, T. Tudrej, F. Ture, Z. Wu, L. Xu, and B. Baldwin. 2025. Non-determinism of "deterministic" llm settings. *Preprint*, arXiv:2408.04667.

J. L. Austin. 1975. *How to Do Things with Words*. Oxford University Press.

M. Avalle, N. Di Marco, G. Etta, E. Sangiorgio, S. Alipour, A. Bonetti, L. Alvisi, A. Scala, A. Baronchelli, M. Cinelli, et al. 2024. Persistent interaction patterns across social media platforms and over time. *Nature*, 628(8008):582–589.

Y. Bang, T. Yu, A. Madotto, Z. Lin, M. Diab, and P. Fung. 2023. Enabling classifiers to make judgements explicitly aligned with human values. In *Proceedings of the 3rd Workshop on Trustworthy NLP*, pages 311–325, Toronto, Canada.

R. Bawden. 2021. Understanding dialogue: Language use and social interaction. *Computational Linguistics*, 47(3):703–705.

T. Beck, H. Schuff, A. Lauscher, and I. Gurevych. 2024. Sensitivity, performance, robustness: Deconstructing the effect of sociodemographic prompting. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2589–2615, Malta.

S. Benesch, D. Ruths, K. P. Dillon, H. M. Saleem, and L. Wright. 2016. Counterspeech on twitter: A field study. dangerous speech project.

J. Bisbee, J. D. Clinton, C. Dorff, B. Kenkel, and J. M. Larson. 2024. Synthetic replacements for human survey data? the perils of large language models. *Political Analysis*, 32(4):401–416.

R. Bose, I. Perera, and B. Dorr. 2023. Detoxifying online discourse: A guided response generation approach for reducing toxicity in user-generated text. In *Proceedings of the First Workshop on Social Influence in Conversations*, pages 9–14, Toronto, Canada.

J. W. Burton, E. Lopez-Lopez, S. Hechtlinger, et al. 2024. How large language models can reshape collective intelligence. *Nature Human Behaviour*, 8:1643–1655.

A. Bächtiger, M. Gerber, and E. Fournier-Tombs. 2022. 83discourse quality index. In *Research Methods in Deliberative Democracy*. Oxford University Press.

A. Bächtiger, S. Niemeyer, M. Neblo, M. R. Steenbergen, and J. Steiner. 2010. Disentangling diversity in deliberative democracy: Competing theories, their blind spots and complementarities. *Journal of Political Philosophy*, 18(1):32–63.

A. Cervone and G. Riccardi. 2020. Is this dialogue coherent? learning from dialogue acts and entities. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 162–174, online. Association for Computational Linguistics.

J. P. Chang and C. Danescu-Niculescu-Mizil. 2019. Trouble on the horizon: Forecasting the derailment of online conversations as they develop. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4743–4754, Hong Kong, China. Association for Computational Linguistics.

G. Chen, L. Cheng, L. A. Tuan, and L. Bing. 2024. Exploring the potential of large language models in computational argumentation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2309–2330.

M. Chen, L. Frermann, and J. H. Lau. 2025. WHoW: A cross-domain approach for analysing conversation moderation. In *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, Albuquerque, New Mexico. Association for Computational Linguistics.

P. Cheng, T. Hu, H. Xu, Z. Zhang, Y. Dai, L. Han, and N. Du. 2024a. Self-playing adversarial language game enhances llm reasoning. *ArXiv*, abs/2404.10642.

P. Cheng, T. Hu, H. Xu, Z. Zhang, Y. Dai, L. Han, and N. Du. 2024b. Self-playing adversarial language game enhances llm reasoning. *ArXiv*, abs/2404.10642.

H. Cho, S. Liu, T. Shi, D. Jain, B. Rizk, Y. Huang, Z. Lu, N. Wen, J. Gratch, E. Ferrara, and J. May. 2024. Can language model moderators improve the health of online discourse? In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7478–7496, Mexico City, Mexico.

G. Cimino, C. Li, G. Carenini, and V. Deufemia. 2024. Coherence-based dialogue discourse structure extraction using open-source large language models. In *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 297–316, Kyoto, Japan.

S. Concannon and M. Tomalin. 2024. Measuring perceived empathy in dialogue systems. *AI & Society*, 39:2233–2247.

S. Cresci, A. Trujillo, and T. Fagni. 2022. Personalized interventions for online moderation. In *Proceedings of the 33rd ACM Conference on Hypertext and Social Media*, page 248–251, New York, NY, USA.

C. Danescu-Niculescu-Mizil, L. Lee, B. Pang, and J. Kleinberg. 2012. Echoes of power: language effects and power differences in social interaction. In *Proceedings of the 21st International Conference on World Wide Web*, page 699–708, New York, NY, USA.

C. De Kock, T. Stafford, and A. Vlachos. 2022. How to disagree well: Investigating the dispute tactics used on Wikipedia. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3824–3837, Abu Dhabi, United Arab Emirates.

C. De Kock and A. Vlachos. 2021. I beg to differ: A study of constructive disagreement in online conversations. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2017–2027, Online.

C. Degeling, S. M. Carter, and L. Rychetnik. 2015. Which public and why deliberate? – a scoping review of public deliberation in public health and health policy research. *Social Science & Medicine*, 131:114–121.

M. El-Assady, A. Hautli-Janisz, V. Gold, M. Butt, K. Holzinger, and D. Keim. 2017. Interactive visual analysis of transcribed multi-party discourse. In *Proceedings of ACL 2017, System Demonstrations*, pages 49–54, Vancouver, Canada.

Cornell eRulemaking Initiative. 2017. Ceri (cornell e-rulemaking) moderator protocol. Cornell e-Rulemaking Initiative Publications, 21.

N. Falk, I. Jundi, E. M. Vecchi, and G. Lapesa. 2021. Predicting moderation of deliberative arguments: Is argument quality the key? In *Proceedings of the 8th Workshop on Argument Mining*, pages 133–141, Punta Cana, Dominican Republic.

N. Falk and G. Lapesa. 2023. Bridging argument quality and deliberative quality annotations with adapters. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 2469–2488.

N. Falk, E. Vecchi, I. Jundi, and G. 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, St. Julian's, Malta.

X. Feng and B. Qin. 2022. A survey on dialogue summarization: Recent advances and new frontiers. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 5453–5460. Survey Track.

A. Ferron, A. Shore, E. Mitra, and A. Agrawal. 2023. MEEP: Is this engaging? prompting large language models for dialogue evaluation in multilingual settings. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2078–2100, Singapore. Association for Computational Linguistics.

O. Ferschke, I. Gurevych, and Y. Chebotar. 2012. Behind the article: Recognizing dialog acts in Wikipedia talk pages. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 777–786, Avignon, France.

J. Fishkin, N. Garg, L. Gelauff, A. Goel, K. Munagala, S. Sakshuwong, A. Siu, and S. Yandamuri. 2018. Deliberative democracy with the online deliberation platform. In *The 7th AAAI Conference on Human Computation and Crowdsourcing (HCOMP 2019). HCOMP*.

E. Fournier-Tombs and M. K. MacKenzie. 2021. Big data and democratic speech: Predicting deliberative quality using machine learning techniques. *Methodological Innovations*, 14(2):20597991211010416.

D. M. Friess. 2018. Letting the faculty deliberate: Analyzing online deliberation in academia using a comprehensive approach. *Journal of Information Technology & Politics*, 15(2):155–177.

S. Furniturewala and K. Jaidka. 2024. Empaths at WASSA 2024 empathy and personality shared task: Turn-level empathy prediction using psychological indicators. In *Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 404–411, Bangkok, Thailand. Association for Computational Linguistics.

R. Gao, M. Chen, L. Frermann, and J. H. Lau. 2025. Moderation matters: Measuring conversational moderation impact in English as a second language group discussion. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 2070–2095, Vienna, Austria. Association for Computational Linguistics.

L. Gelauff, L. Nikolenko, S. Sakshuwong, J. Fishkin, A. Goel, K. Munagala, and A. Siu. 2023. *Achieving parity with human moderators*, pages 202–221. Routledge.

M. H. Gelula. 1997. Clinical discussion sessions and small groups. *Surgical Neurology*, 47(4):399–402.

M. Gerber, A. Bächtiger, S. Shikano, S. Reber, and S. Rohr. 2018. Deliberative abilities and influence in a transnational deliberative poll (europolis). *British Journal of Political Science*, 48(4):1093–1118.

H. Gimpel, S.n Lahmer, M. Wöhl, et al. 2024. Digital facilitation of group work to gain predictable performance. *Group Decision and Negotiation*, 33:113–145.

K. Gligori'c, T. Zrnic, C. Lee, E. J. Candes, and D. Jurafsky. 2024. Can unconfident llm annotations be used for confident conclusions? *ArXiv*, abs/2408.15204.

J. I. Goñi. 2024. What is "dialogue" in public engagement with science and technology? bridging sts and deliberative democracy. *Minerva*.

P. Graham. 2008. How to disagree. Accessed: 2024-06-24.

T. Graham and T Witschge. 2003. In search of online deliberation: Towards a new method for examining the quality of online discussions. *Communications*, 28(2):173–204.

H. P. Grice. 1975. Logic and conversation. *Syntax and semantics*, 3.

M. Habibi, D. Hovy, and C. Schwarz. 2024. The content moderator's dilemma: Removal of toxic content and distortions to online discourse. *Preprint*, arXiv:2412.16114.

J. Hessel and L. Lee. 2019. Something's brewing! early prediction of controversy-causing posts from discussion features. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1648–1659, Minneapolis, Minnesota. Association for Computational Linguistics.

L. Hewitt, A. Ashokkumar, I. Ghezae, and R. Willer. 2024. Predicting results of social science experiments using large language models. Equal contribution, order randomized.

E. Hoque and G. Carenini. 2016. Multiconvis: A visual text analytics system for exploring a collection of online conversations. In *Proceedings of the 21st International Conference on Intelligent User Interfaces*, IUI '16, page 96–107, New York, NY, USA.

J. Hu, S. Floyd, O. Jouravlev, E. Fedorenko, and E. Gibson. 2023. A fine-grained comparison of pragmatic language understanding in humans and language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4194–4213, Toronto, Canada.

Y. Hua, C. Danescu-Niculescu-Mizil, D. Taraborelli, N. Thain, J. Sorensen, and L. Dixon. 2018. WikiConv: A corpus of the complete conversational history of a large online collaborative community. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2818–2823, Brussels, Belgium.

A. Irani, M. Faloutsos, and K. Esterling. 2024. Argusense: Argument-centric analysis of online discourse. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 18, pages 663–675.

B. J. Jansen, S. Jung, and J. Salminen. 2023. Employing large language models in survey research. *Natural Language Processing Journal*, 4:100020.

H. Jin, Y. Zhang, D. Meng, J. Wang, and J. Tan. 2024. A comprehensive survey on process-oriented automatic text summarization with exploration of llm-based methods. *arXiv preprint*.

P. Jwalapuram. 2017. Evaluating dialogs based on Grice's maxims. In *Proceedings of the Student Research Workshop Associated with RANLP 2017*, pages 17–24, Varna.

S. Kaner, Le. Lind, C. Toldi, S. Fisk, and D. Berger. 2007. *Facilitator's Guide to Participatory Decision-Making*. John Wiley & Sons/Jossey-Bass, San Francisco.

H. Kang and T. Qian. 2024. Implanting LLM's knowledge via reading comprehension tree for toxicity detection. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 947–962, Bangkok, Thailand and virtual meeting.

G. Karadzhov, T. Stafford, and A. Vlachos. 2021. Delidata: A dataset for deliberation in multi-party problem solving. *Proceedings of the ACM on Human-Computer Interaction*, 7:1 – 25.

B. Khalid and S. Lee. 2022. Explaining dialogue evaluation metrics using adversarial behavioral analysis. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5871–5883, Seattle, United States. Association for Computational Linguistics.

R. Kies. 2022. Online deliberative matrix. In *Research Methods in Deliberative Democracy*, pages 148–162. Oxford University Press.

S. Kim, J. Eun, J. Seering, and J. Lee. 2021. Moderator chatbot for deliberative discussion: Effects of discussion structure and discussant facilitation. *Proc. ACM Hum.-Comput. Interact.*, 5(CSCW1).

D. Kumar, Y. A. AbuHashem, and Z. Durumeric. 2024. Watch your language: Investigating content moderation with large language models. *Proceedings of the International AAAI Conference on Web and Social Media*, 18(1):865–878.

.C Lambert, F. Choi, and E. Chandrasekharan. 2024. "positive reinforcement helps breed positive behavior": Moderator perspectives on encouraging desirable behavior. *Proc. ACM Hum.-Comput. Interact.*, 8(CSCW2).

R. Langevin, R. J Lordon, T. Avrahami, B. R. Cowan, T. Hirsch, and G. Hsieh. 2021. Heuristic evaluation of conversational agents. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, New York, NY, USA.

J. Lawrence, J. Park, K. Budzynska, C. Cardie, B. Konat, and C. Reed. 2017. Using argumentative structure to interpret debates in online deliberative democracy and erulemaking. *ACM Trans. Internet Technol.*, 17(3).

J. Lawrence and C. Reed. 2020. Argument mining: A survey. *Computational Linguistics*, 45(4):765–818.

Z. Li, J. Zhang, Z. Fei, Y. Feng, and J. Zhou. 2021. Conversations are not flat: Modeling the dynamic information flow across dialogue utterances. 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 128–138, Online.

S.C.R. Lim, W. Cheung, and K. Hew. 2011. Critical thinking in asynchronous online discussion: An investigation of student facilitation techniques. *New Horizons in Education*, 59:52–65.

M. Lipman. 2003. *Thinking in Education*, 2 edition. Cambridge University Press.

Y. Liu, S. Ultes, W. Minker, and W. Maier. 2023. Unified conversational models with system-initiated transitions between chit-chat and task-oriented dialogues. In *Proceedings of the 5th International Conference on Conversational User Interfaces*, CUI '23, New York, NY, USA.

J. Lo and P. McAvoy. 2023. *Debate and Deliberation in Democratic Education*, page 298–310. Cambridge Handbooks in Education. Cambridge University Press.

L. Lugini, C. Olshefski, R. Singh, D. Litman, and A. Godley. 2020. Discussion tracker: Supporting teacher learning about students' collaborative argumentation in high school classrooms. In *Proceedings of the 28th International Conference on Computational Linguistics: System Demonstrations*, pages 53–58, Barcelona, Spain (Online). International Committee on Computational Linguistics (ICCL).

F. Macagno, C. Rapanta, E. Mayweg-Paus, and M. Garcia-Milà. 2022. Coding empathy in dialogue. *Journal of Pragmatics*, 192:116–132.

N. Mansour. 2024. Students' and facilitators' experiences with synchronous and asynchronous online dialogic discussions and e-facilitation in understanding the nature of science. *Education and Information Technologies*, 29:15965–15997.

A. Martinenghi, G. Donabauer, S. Amenta, S. Bursic, M. Giudici, U. Kruschwitz, F. Garzotto, and D. Ognibene. 2024. LLMs of catan: Exploring pragmatic capabilities of generative chatbots through prediction and classification of dialogue acts in boardgames' multi-party dialogues. In *Proceedings of the 10th Workshop on Games and Natural Language Processing @ LREC-COLING 2024*, pages 107–118, Torino, Italia.

B. Mathew, R. Dutt, P. Goyal, and A. Mukherjee. 2019. Spread of hate speech in online social media. In *Proceedings of the 10th ACM Conference on Web Science*, page 173–182, New York, NY, USA.

J. Mendonca, I. Trancoso, and A. Lavie. 2024. ECoh: Turn-level coherence evaluation for multilingual dialogues. In *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 516–532, Kyoto, Japan.

N. Mirzakhmedova, M. Gohsen, C. H. Chang, and B. Stein. 2024. Are large language models reliable argument quality annotators? In *Conference on Advances in Robust Argumentation Machines*, pages 129–146. Springer.

MIT Center for Constructive Communication. 2024. Unpublished training materials developed by the mit center for constructive communication. Guide given to human facilitators.

M.D. Molina and S.S. Sundar. 2022. When AI moderates online content: effects of human collaboration and interactive transparency on user trust. *Journal of Computer-Mediated Communication*, 27(4).

Y. Nam, H. Chung, and U. Hong. 2023. Language artificial intelligences' communicative performance quantified through the gricean conversation theory. *Cyberpsychology, Behavior, and Social Networking*, 26(12):919–923. PMID: 37976199.

T. Neumann, M. De-Arteaga, and S. Fazelpour. 2025. Should you use llms to simulate opinions? quality checks for early-stage deliberation. *Preprint*, arXiv:2504.08954.

E.W.T. Ngai, M.C.M. Lee, M. Luo, P.S.L. Chan, and T. Liang. 2021. An intelligent knowledge-based chatbot for customer service. *Electronic Commerce Research and Applications*, 50:101098.

V. Niculae and C. Danescu-Niculescu-Mizil. 2016. Conversational markers of constructive discussions. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 568–578, San Diego, California.

J. Park, S. Klingel, C. Cardie, M. Newhart, C. Farina, and J.J. Vallbé. 2012. Facilitative moderation for online participation in erulemaking. In *Proceedings of the 13th Annual International Conference on Digital Government Research*, page 173–182, New York, NY, USA.

J.S. Park, J.C. O'Brien, C.J. Cai, M.R. Morris, P. Liang, and M.S. Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*.

J.S. Park, L. Popowski, C.J. Cai, M.R. Morris, P. Liang, and M.S. Bernstein. 2022. Social simulacra: Creating populated prototypes for social computing systems. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*, UIST '22, New York, NY, USA.

F. Pradel, J. Zilinsky, S. Kosmidis, and Y. Theocharis. 2024. Toxic speech and limited demand for content moderation on social media. *American Political Science Review*, 118(4):1895–1912.

N. Raj Prabhu, C. Raman, and H. Hung. 2021. Defining and quantifying conversation quality in spontaneous interactions. In *Companion Publication of the 2020 International Conference on Multimodal Interaction*, ICMI '20 Companion, page 196–205, New York, NY, USA.

P. Rescala, M.H. Ribeiro, T. Hu, and R. West. 2024. Can language models recognize convincing arguments? In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 8826–8837, Miami, Florida, USA.

H. Rheingold. 2000. *The Virtual Community: Homesteading on the Electronic Frontier*. The MIT Press.

R. Rose-Redwood, R. Kitchin, L. Rickards, U. Rossi, A. Datta, and J. Crampton. 2018. The possibilities and limits to dialogue. *Dialogues in Human Geography*, 8(2):109–123.

L. Rossi, K. Harrison, and I. Shklovski. 2024. The problems of llm-generated data in social science research. *Sociologica*, 18(2):145–168.

L. Ruis, A. Khan, S. Biderman, S. Hooker, T. Rocktäschel, and E. Grefenstette. 2023. The goldilocks of pragmatic understanding: fine-tuning strategy matters for implicature resolution by llms. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA.

U. Russmann and A. Lane. 2016. Discussion. dialogue, and discoursel doing the talk: Discussion, dialogue, and discourse in action — introduction. *International Journal of Communication*, 10.

M. Saeidi, M. Yazdani, and A. Vlachos. 2021. Cross-policy compliance detection via question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8622–8632, Online and Punta Cana, Dominican Republic.

H. Sak, A. W. Senior, and F. Beaufays. 2014. Long short-term memory based recurrent neural network architectures for large vocabulary speech recognition. *ArXiv*, abs/1402.1128.

M. Sap, S. Gabriel, L. Qin, D. Jurafsky, N. A. Smith, and Y. Choi. 2020. Social bias frames: Reasoning about social and power implications of language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5477–5490, Online.

B. Schaffner, A. N. Bhagoji, S. Cheng, J. Mei, J.L. Shen, G. Wang, M. Chetty, N. Feamster, G. Lakier, and C. Tan. 2024. "Community guidelines make this the best party on the internet": An in-depth study of online platforms' content moderation policies. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA.

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

H. Schroeder, D. Roy, and J. 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*, pages 13985–14001, Bangkok, Thailand.

J. R. Searle. 1969. *Speech Acts: An Essay in the Philosophy of Language*. Cambridge University Press.

A. See, S. Roller, D. Kiela, and J. Weston. 2019. What makes a good conversation? how controllable attributes affect human judgments. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1702–1723, Minneapolis, Minnesota. Association for Computational Linguistics.

J. Seering. 2020. Reconsidering self-moderation: the role of research in supporting community-based models for online content moderation. *Proc. ACM Hum.-Comput. Interact.*, 4(CSCW2).

F. Shahid, M. Dittgen, M. Naaman, and A. Vashistha. 2024. Examining human-AI collaboration for co-writing constructive comments online. *arXiv preprint arXiv:2411.03295*.

X. Shi, J. Liu, and Y. Song. 2024. BERT and LLM-based multivariate hate speech detection on twitter: Comparative analysis and superior performance. In *Artificial Intelligence and Machine Learning*, pages 85–97, Singapore. Springer Nature Singapore.

E. Sjølie, A. Strømme, and J. Boks-Vlemmix. 2021. Team-skills training and real-time facilitation as a means for developing student teachers' learning of collaboration. *Teaching and Teacher Education*, 107:103477.

C.T. Small, I. Vendrov, E. Durmus, H. Homaei, E. Barry, J. Cornebise, T. Suzman, D. Ganguli, and C. Megill. 2023. Opportunities and risks of LLMs for scalable deliberation with Polis. *ArXiv*, abs/2306.11932.

E. Smith, O. Hsu, R. Qian, S. Roller, Y-L. Boureau, and J. Weston. 2022. Human evaluation of conversations is an open problem: Comparing the sensitivity of various methods for evaluating dialogue agents. In *Proceedings of the 4th Workshop on NLP for Conversational AI*, pages 77–97, Dublin, Ireland. Association for Computational Linguistics.

S. Sravanthi, M. Doshi, P. Tankala, R. Murthy, R. Dabre, and P. Bhattacharyya. 2024. PUB: A pragmatics understanding benchmark for assessing LLMs' pragmatics capabilities. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12075–12097, Bangkok, Thailand.

M. Steenbergen, A. Bächtiger, M. Spörndli, and J. Steiner. 2003. Measuring political deliberation: A discourse quality index. *Comparative European Politics*, 1:21–48.

A. Stolcke, K. Ries, N. Coccaro, E. Shriberg, R. Bates, D. Jurafsky, P. Taylor, R. Martin, C. Van Ess-Dykema, and M. Meteer. 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational Linguistics*, 26(3):339–374.

J. Stromer-Galley. 2007. Measuring deliberation's content: A coding scheme. *Journal of Deliberative Democracy*, 3(1):25–44.

K. Sun, S. Moon, P. Crook, S. Roller, B. Silvert, B. Liu, Z. Wang, H. Liu, E. Cho, and C. Cardie. 2021. Adding chit-chat to enhance task-oriented dialogues. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1570–1583, Online. Association for Computational Linguistics.

C. Tan, V. Niculae, C. Danescu-Niculescu-Mizil, and L. Lee. 2016. Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions. In *Proceedings of the 25th International Conference on World Wide Web*, WWW '16, page 613–624, Republic and Canton of Geneva, CHE.

A. Taubenfeld, Y. Dover, R. Reichart, and A. Goldstein. 2024. Systematic biases in llm simulations of debates. *ArXiv*, abs/2402.04049.

M.H. Tessler, M.A. Bakker, D. Jarrett, H. Sheahan, M.J. Chadwick, R. Koster, G. Evans, L. Campbell-Gillingham, T.Collins, D.C. Parkes, M. Botvinick, and C. Summerfield. 2024. AI can help humans find common ground in democratic deliberation. *Science*, 386(6719).

The Commons. 2025. The commons project. Accessed: 2025-01-27.

M. Trenel. 2009. Facilitation and inclusive deliberation. In *Online Deliberation: Design, Research, and Practice*, pages 253–257. CSLI Publications/University of Chicago Press.

A. Trujillo and S. Cresci. 2022. Make reddit great again: Assessing community effects of moderation interventions on r/the_donald. *Proceedings of the ACM on Human-Computer Interaction*, 6:1 – 28.

L. L. Tsai, A. Pentland, A. Braley, N. Chen, J. R. Enríquez, and A. Reuel. 2024. Generative AI for Pro-Democracy Platforms. *An MIT Exploration of Generative AI*. Https://mit-genai.pubpub.org/pub/mn45hexw.

J.A. Tucker, A. Guess, P. Barberá, C. Vaccari, A. Siegel, S. Sanovich, D. Stukal, and B. Nyhan. 2018. Social media, political polarization, and political disinformation: A review of the scientific literature. *SSRN Electronic Journal*.

D. Ulmer, E. Mansimov, K. Lin, L. Sun, X. Gao, and Y. Zhang. 2024. Bootstrapping LLM-based task-oriented dialogue agents via self-talk. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 9500–9522, Bangkok, Thailand.

E.M. Vecchi, N. Falk, I. Jundi, and G. Lapesa. 2021. Towards argument mining for social good: A survey. In *Proceedings of the 59th Annual Meeting of ACL and 11th International Joint Conference on NLP*, pages 1338–1352, Online.

A. Veglis. 2014. Moderation techniques for social media content. In *Social Computing and Social Media*, pages 137–148, Cham. Springer International Publishing.

M. Verkuyl, E. Violato, T. Southam, M. Lavoie-Tremblay, S. Goldsworthy, D. MacEachern, and L. Atack. 2024. Facilitators' experiences with virtual simulation and their impact on learning. *Advances in Simulation*, 9(1):54.

H. Wachsmuth, G. Lapesa, E. Cabrio, A. Lauscher, J. Park, E.M. Vecchi, S. Villata, and T. Ziegenbein. 2024. Argument quality assessment in the age of instruction-following large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation*, pages 1519–1538, Torino, Italia.

H. Wachsmuth, N. Naderi, Y. Hou, Y. Bilu, V. Prabhakaran, T.A. Thijm, G. Hirst, and B. Stein. 2017. Computational argumentation quality assessment in natural language. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 176–187.

M. Walker, J. F. Tree, P. Anand, R. Abbott, and J. King. 2012. A corpus for research on deliberation and debate. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 812–817, Istanbul, Turkey.

Y. Wang and Y. T. Chang. 2022. Toxicity detection with generative prompt-based inference. *ArXiv*, abs/2205.12390.

Y. Wang, X. Chen, B. He, and L. Sun. 2023. Contextual interaction for argument post quality assessment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10420–10432.

M. Warner, A. Strohmayer, M. Higgs, and L. Coventry. 2025. A critical reflection on the use of toxicity detection algorithms in proactive content moderation systems. *International Journal of Human-Computer Studies*, 198:103468.

K. White, N. Hunter, and K. Greaves. 2024. *facilitating deliberation - a practical guide*. Mosaic Lab.

T. P. Wilson, J. M. Wiemann, and D. H. Zimmerman. 1984. Models of turn taking in conversational interaction. *Journal of Language and Social Psychology*, 3(3):159–183.

Z. Xu and J. Jiang. 2024. Multi-dimensional evaluation of empathetic dialogue responses. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 2066–2087, Miami, Florida, USA. Association for Computational Linguistics.

Y. Yeh, M. Eskenazi, and S. Mehri. 2021. A comprehensive assessment of dialog evaluation metrics. In *The First Workshop on Evaluations and Assessments of Neural Conversation Systems*, pages 15–33, Online. Association for Computational Linguistics.

D. Yu, L. Li, H. Su, and M. Fuoli. 2024. Assessing the potential of llm-assisted annotation for corpus-based pragmatics and discourse analysis. *International Journal of Corpus Linguistics*, 29(4):534–561.

A. Zhang, B. Culbertson, and P. Paritosh. 2017. Characterizing online discussion using coarse discourse sequences. *Proceedings of the International AAAI Conference on Web and Social Media*, 11(1):357–366.

C. Zhang, L. D'Haro, C. Tang, K. Shi, G. Tang, and H. Li. 2023. xDial-eval: A multilingual open-domain dialogue evaluation benchmark. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5579–5601, Singapore.

C. Zhang, L. F. D'Haro, Y. Chen, M. Zhang, and H. Li. 2024. A comprehensive analysis of the effectiveness of large language models as automatic dialogue evaluators. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19515–19524.

J. Zhang, J. Chang, C. Danescu-Niculescu-Mizil, L. Dixon, Y. Hua, D. Taraborelli, and N. Thain. 2018. Conversations gone awry: Detecting early signs of conversational failure. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1350–1361, Melbourne, Australia.

L. Zhou, Y. Farag, and A. Vlachos. 2024. An LLM feature-based framework for dialogue constructiveness assessment. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5389–5409, Miami, Florida, USA. Association for Computational Linguistics.

C. Ziems, W. Held, O. Shaikh, J. Chen, Z. Zhang, and D. Yang. 2024. Can large language models transform computational social science? *Computational Linguistics*, 50(1):237–291.

A Acronyms

NLP Natural Language Processing

ML Machine Learning

LLM Large Language Model

AM Argument Mining

ML Machine Learning

IR Information Retrieval

AQ Argument Quality

ESL English as a Second Language

B Keywords for Literature Query

Keyword Selection
online discussions, deliberation, dialogue, discussion evaluation, discussion metrics, dialogue, deliberation, NLP, AI, discussion quality, argument mining, survey, LLM, conversation, moderation, facilitation, communication, democracy AI dialogue systems, group dynamics

Table 2: Keywords for search engine queries

C Terminology Background

Here, we explain our reasoning for choosing and disambiguating certain terms (see §2). The definitions of the terms can be found in Table 3.

Facilitation vs. Moderation "Moderation", as a term, is more common in Computer Science and NLP, while facilitation is prevalent in Social Sciences (Vecchi et al., 2021; Kaner et al., 2007; Trenel, 2009). Moderators enforce rules and ensure orderly interactions, usually with the threat of disciplinary action, though they can also act as community leaders (Falk et al., 2024; Seering, 2020; eRulemaking Initiative, 2017). Facilitators, on the other hand, guide discussions, promote participation and structured dialogue, particularly in online deliberation and education platforms (Asterhan and Schwarz, 2010). Despite these distinctions, the terms are sometimes used interchangeably (Cho et al., 2024; Park et al., 2012; Kim et al., 2021), while it is also common for moderators to use facilitation tactics (eRulemaking Initiative, 2017; Park et al., 2012; Kim et al., 2021; Cho et al., 2024; Schluger et al., 2022).

Pre-moderation and Post-moderation Multiple taxonomies have been proposed to describe the temporal dimension of moderation; that is, when moderator action is applied in relation to when the content is visible to the users (Veglis, 2014; Schluger et al., 2022). These taxonomies are very similar to each other, and usually boil down to the following distinctions:

- *Pre-moderation:* The user is dissuaded, or prevented from, posting harmful content. Pre-moderation techniques can include nudges at the writing stage (Argyle et al., 2023), reminders about platform rules (Schluger et al., 2022), or even a moderation queue where posts have to be approved before being visible to others (Schluger et al., 2022).
- *Real-Time:* The moderator is part of the discussion and intervenes like a referee would during a match.
- *Ex-post:* The moderator is called after a possible incident has been flagged and makes the final call.

Discussion, Deliberation, Dialogue, Debate There is little to no consensus on how to properly define terms such as "discussion" and "dialogue" (Russmann and Lane, 2016; Goñi, 2024). In this section, we attempt to disambiguate the use of such terms for the purposes of our survey and based on the existing related work. First, our study focuses on **discussions**, a broader term encompassing various informal and formal exchanges, including online discussions in fora (Russmann and Lane, 2016), with which we are mainly concerned. In contrast, **dialogue** refers to collaborative interactions in which participants work toward a shared understanding and alignment (Rose-Redwood et al., 2018; Bawden, 2021; Goñi, 2024). Studies on dialogue emphasize its cooperative nature, aiming for mutual insight rather than competition (Bawden, 2021). Dialogue can also refer to dialogue systems, a major NLP sub-area, traditionally including both task-oriented dialogues and casual conversation (Eliza-like)[3] "chatbots" (Liu et al., 2023; Sun et al., 2021).

[3] `http://web.njit.edu/~ronkowit/eliza.html`

A more specific concept is **deliberation**, which involves structured discussions aimed at informed decision-making, often prioritizing reasoned argumentation and the consideration of diverse perspectives (Degeling et al., 2015; Lo and McAvoy, 2023). Meanwhile, **debate** is typically adversarial, where participants focus on persuading others or defending their positions. Unlike dialogue or deliberation, debate centers more on winning or convincing, making it less about collective reasoning and more about rhetorical effectiveness (Lo and McAvoy, 2023). Debates also typically have much stricter (and enforced) rules than other discussions.

For this study, we specifically focus on online written discussions, particularly those occurring in **thread- or tree-style** formats (Seering, 2020). A thread is a collection of messages or posts grouped together in an online forum, discussion board, or messaging platform (such as Reddit). It begins with an initial post (often called the original post, or OP), and subsequent replies are ordered either chronologically or by relevance. Threads usually address a specific topic or question and allow users to engage in discussions about that subject. A thread may grow as users contribute more responses. It must be noted, however, that this type of discussion can contain elements from all the other discussion styles. For example, the adversarial element of the debates, or the argumentative element that can be found both in dialogues and deliberations.

Discussion Quality The success of a discussion is often subjective, influenced by a variety of factors such as the cultural background and linguistic proficiency of the participants (Zhang et al., 2018), as well as their level of engagement (See et al., 2019). It also depends on the type of the discussion, since some types of discussions, such as deliberations or debates, may not aim at consensus. Given these complexities, we adopt the definition proposed by Raj Prabhu et al. (2021), which views the perceived *discussion quality* as a measurement that attempts to quantify interactions by taking into account multiple socio-dimensional aspects of individual experiences and abilities.

D Methodology

The search and article selection of this survey was conducted using specific keywords in academic search engines (e.g., Google Scholar, Semantic Scholar, Scopus), digital libraries and repositories

Concept	Definition and Characteristics
Discussion	Broad term encompassing informal and formal exchanges, including online discussions in fora. Can involve elements of debate, dialogue, and deliberation.
Dialogue	Collaborative interaction aimed at shared understanding and alignment. Emphasizes cooperation rather than competition. Also refers to dialogue systems in NLP (task-oriented or chatbot conversations).
Deliberation	Structured discussion focusing on informed decision-making with reasoned argumentation and diverse perspectives. Less about persuasion, more about collective reasoning.
Debate	Adversarial interaction where participants aim to persuade or defend positions rather than achieve mutual understanding. Focused on rhetorical effectiveness.
Thread-style Discussions	Online discussions structured in tree/thread formats (e.g., Reddit). Can incorporate elements of all rhetorical styles (debate, dialogue, deliberation).
Discussion Quality	Subjective measure influenced by cultural background, engagement, and type of discussion. Defined by socio-dimensional aspects of participant experiences.
Moderation	Ensures orderly interactions by enforcing guidelines. Moderators can be volunteers or employees, often associated with disciplinary actions.
Facilitation	Encourages equal participation and organizes discussion flow. More common in deliberative and educational contexts, though often used interchangeably with moderation.

Table 3: Definition of terms used in this survey.

(e.g., ACL Anthology, ACM Digital Library, IEEE Xplore, JSTOR). We focused on peer-reviewed publications written in English between 2014 and 2024, granting exceptions only for established works predating this period. Additionally, we reviewed other cited papers that appeared highly relevant, provided they were peer-reviewed and cited by more than 20 citations of other researchers, unless the topic was very niche, in which case we judged by its content. The search strategy incorporated keywords and phrases related to LLMs, discussion facilitation, and discussion evaluation. The list of keywords used is provided in Table 2. The search was further informed by existing survey articles, such as those by Vecchi et al. (2021) and Wachsmuth et al. (2024), which served as starting points both for identifying relevant literature and for specifying the vocabulary used in the keyword search.

E Discussion Quality Taxonomy

In this part of the Appendix, we present a table summarizing the discussion evaluation taxonomy (§4). The dimensions are outlined alongside both pre-LLM and LLM-based approaches, while also highlighting their respective contributions to facilitation. The dimensions are color-coded for clarity, with orange indicating associated dimensions that could serve as early signs of potential derailment, green marking signs of constructive growth—i.e., conversations going well or worth participating in—and pink denoting interaction dynamics.

F Online Discussion Example with Color-coded Politeness Markers

Table 5 highlights key politeness-related linguistic features such as hedging, personal references, sentiment, and direct questions. These features are essential in the context of facilitation, where the goal is to guide conversations constructively, maintain safety, and foster mutual understanding. By identifying these elements, the facilitator (human or automatic) can better interpret the tone, intent, and emotional weight of each utterance. For example, detecting hedging or positive sentiment can guide the model to adopt a more collaborative tone, while recognizing negative sentiment or accusatory second-person references may prompt it to de-escalate tension and encourage constructive dialogue.

Dimension	**Facilitation Use**	**Pre-LLM Approaches**	**LLM Approaches**
Structure & Logic			
Argument structure & analysis	Spot claim-evidence chains; raise early-warning flags; keep debate fact-centred	Argument-mining pipelines: claim/premise detection; AQ scoring; graph & neural models	Zero/few-shot AQ labelling; argument-structure parsing; on-the-fly argument-map summaries
Coherence & flow	Detect topic drift; redirect or bridge gaps	Entity-grid & sequential coherence models; topic modelling; dialogue state tracking	Prompted coherence scoring; chain-of-thought flow checks; off-topic suggestions
Turn-taking	Monitor balance (entropy/Gini); nudge silent voices; avoid dominance	Turn-entropy / Gini metrics; rule-based alarms	Context-window turn counts; balanced-participation prompts
Linguistic markers	Track hedges, 2nd-person spikes, jargon; trigger clarification or civility nudges	Lexicon features; n-gram-based hedging detectors	Style-transfer rephrasers; embedding hedge detection; tone-repair suggestions
Speech & dialogue acts	Identify interruptions, proposals, question types; score deliberative quality	Dialogue-act tagging with ISO/DAMSL labels	Few-shot Dialogue Act tagging; tactic selection based on Dialogue Act patterns
Pragmatic comprehension	Resolve implicatures & sarcasm; surface hidden misunderstandings	Commonsense reasoning (Knowledge Base + neural); limited coverage	In-context reasoning; auto clarifying questions
Social Dynamics			
Politeness	Forecast derailment; issue civility nudges or positive reinforcement	Politeness lexicons; domain-independent classifiers	Annotation & polite rewrites; policy-violation explanations
Power & status	Detect dominance; invite low-status voices; rebalance floor	Style-matching, pronoun analysis; social-role features	Power imbalance estimation; moderator suggestions
Disagreement	Distinguish constructive vs destructive dissent; de-escalate	Graham-hierarchy / stance detection	Few-shot labelling; automatic reframing prompts
Emotion & Behavior			
Empathy	Encourage empathic turns; highlight emotional cues	Lexicon/coding empathy classifiers; affective features	Perceived-empathy scoring; supportive paraphrases
Toxicity	Flag harmful language; decide moderation step	BERT/toxicity classifiers; detox lexicons	Detection + rewrite suggestions; policy chat
Sentiment	Track emotional climate; intervene at negativity spikes	Lexicon & neural sentiment analysis	Prompt-based labelling; tone-shift detection
Controversy	Sense polarization; invite balancing views	Topic-polarity metrics; ideology models	Ideology tagging; polarity-aware summaries
Constructiveness	Stream score; escalate or summarize based on trend	Feature-based classifiers (linguistic, discourse)	Constructive-rewrite coaching
Engagement & Impact			
Engagement	Detect lulls or dominance; prompt interaction	Turn/word counts; reply-time gaps	Auto-recaps; invite quiet users
Persuasion	Spotlight evidence-based arguments; dampen manipulation	Lexical overlap; ethos/pathos/logos; persuasion prediction	Outcome prediction; neutral framing suggestions
Diversity & Informativeness	Monitor viewpoint spread & info density	Topic-diversity indices; IR-based scoring	Simulate perspectives; propose links

Table 4: Summary of discussion quality dimensions and corresponding pre-LLM and LLM-based facilitation strategies.

Turn	Utterance
0	Why should we help people based on race, and say "we'll help everyone who's black, because they could be poor" instead of just "we'll help everyone who's poor, in which black people make up a proportionally larger amount"?
1	That study is worse than useless unless it also distinguishes between "black sounding" names that are associated with wealth and poverty.
2	That wouldn't discount it, that would just add another intersectional axis to investigate. >which I know without looking that it didn't. How rational.
3	It's certainly more rational than unquestioningly swallowing everything I read, as some people do. Did this study of yours also test difficult to pronounce Polish names, or Russian names? Or would that have interfered too much with the foregone conclusion they were attempting to reach?
4	Are you implying that's what I have done? You may be the only one making assumptions here.

Table 5: Dissucssion example from the Reddit Change My View dataset (Chang and Danescu-Niculescu-Mizil, 2019). Color indicates politeness-related features: hedging, 1st person reference, 2nd person reference, direct questions, negative sentiment and positive sentiment. The annotation was produced with a soon-to-be-released annotation toolkit for discussion evaluation.