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

Anonymous ACL submission

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 facilitation agents that 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, along with a new taxonomy on conversation facilitation datasets, (c) an LLM-oriented roadmap of good practices and future research directions, from technological and societal perspectives.

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/learning (Mansour, 2024) or clinical discussions (Gelula, 1997), without much research conducted for thread-like discussions.

Improving the quality of online discussions presupposes being able to define and measure *discussion quality*. Here, work from Social Science, but also Argument Mining (AM) (Lapesa et al., 2024), can again provide several ideas on dimensions (aspects) of discussion quality, such as logical cohesion and constructiveness, as well as ideas on methods to measure quality along each dimension.

This paper surveys research from Social Science and relevant NLP areas (e.g., AM, Sentiment Analysis, Toxicity Detection), focusing on how Large Language Models (LLMs) can facilitate human discussions—similar to human facilitators (Kahane, 2013). While prior 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), their connection to Social Science remains underexplored. In this survey, we include methods from Social Science (e.g., linguistics) when discussing approaches for evaluating online discussions, as well as when exploring intervention strategies (e.g., facilitative tactics). LLM-based facilitation again presupposes defining and evaluating discussion quality. This is even more necessary in the case of LLMs, because of their rapid deployment, potential biases, and long-term societal consequences. Without continuous and thorough assessment, we risk implementing LLM-based facilitation systems that may be ineffective, biased, or even harmful, before their full implications are understood.

Therefore, we survey discussion evaluation aspects and their feasibility with LLMs, introducing a new taxonomy inspired by deliberation studies. We map tasks suited for ML models, LLMs, and humans, aggregate multidimensional insights on facilitation strategies, and outline future capabil-

ities of LLMs. Additionally, we compare major datasets and introduce a taxonomy of tasks relevant to facilitation. Our work focuses mostly on written thread-like discussions (e.g., Reddit threads).

Our findings show that (a) many discussion evaluation aspects coexist in literature; we propose a precise taxonomy to alleviate this, (b) advancements in machine and machine-enhanced human intervention show significant promise in improving the quality and timeliness of facilitation methods; (c) there is a general lack of datasets for studying facilitation; we aggregate existing datasets. We posit that LLM-generated discussions, could become an asset to test facilitation strategies in diverse discussions.

2 Terminology

Given the numerous aspects to consider regarding discussion quality and facilitation, we narrow the scope of the survey and clarify the terminology we use. We highly recommend that the reader refer to 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 We find that the terms 'moderation' and 'facilitation' usually refer to the same process in the literature, with 'facilitation' being used more in Social Science, and 'moderation' in NLP (Vecchi et al., 2021). Thus, we will be using them interchangeably in this survey.

Ex-Post moderation This survey mainly focuses on 'Real-Time, Ex-Post-moderation', i.e., moderation happening 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/facilitator 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 collaborative nature of dialogues (Rose-Redwood et al., 2018; Bawden, 2021; Goñi, 2024) and the competitive nature of debates (Lo and McAvoy, 2023).

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

3 Comparison to Other Surveys

Few surveys have considered discussion quality evaluation and facilitation from a perspective that encompasses ideas from both Social Science and Computer Science, as the present survey does. The closest related surveys are those by Wachsmuth et al. (2024) and Vecchi et al. (2021). Wachsmuth et al. (2024) focus primarily on discussion *evaluation* on its own, rather than exploring how it relates to facilitation, which is one of the main goals of our survey. Furthermore, their survey predominantly centers on concepts from AM. On the other hand, the survey of Vecchi et al. (2021) is rooted largely in AM research. 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, convergence, and problem-solving. In other words, they argue that advancing AM for social good requires a collaborative effort between AM and Social Science. Building on this notion, our survey focuses on both discussion evaluation and facilitation, incorporating ideas from Social Science into NLP-based approaches.

4 Survey 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 (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 is very niche, in which case we judge by its content. The search strategy incorporated keywords and phrases related to LLMs, discussion facilitation, and discussion evaluation. The list of key-

words used is provided in Appendix B 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.

5 Discussion Quality Evaluation

According to Kies (2022), deliberation quality is assessed along three dimensions: (1) *Deliberative Presence*, examining inclusion (ability of all interested parties to participate in the deliberation process), discursive equality (equal opportunities of articulation), justification (well-reasoned arguments), and reciprocity (participants listen and react to each other); (2) *Deliberative Attitudes*, covering reflexivity (being open-minded to be convinced by the arguments of others), empathy (shared understanding towards other views and opinions), and sincerity (disclose true positions and beliefs); (3) *Deliberative Outcomes*, assessing plurality (range of opinions) and external impact on broader political discourse. However, the focus of this survey is online written discussions that do not focus exclusively on the deliberative goal of discourse (Gerber et al., 2018), unlike the work of Kies (2022).

For that reason, we also take into account Social Science, which offers different theories and definitions about Argument Quality (AQ), and which focuses not only on the argument itself but also on the interaction and social dynamics between participants (Falk and Lapesa, 2023). In particular, we leverage Social Science (particularly Deliberative Theory (DT)) quality notions (Bächtiger et al., 2022, 2010; Steenbergen et al., 2003), and discussion quality aspects (Falk and Lapesa, 2023). We expand the work of Kies (2022), defining a new broader taxonomy for the employed discussion quality aspects.

5.1 Structure and Logic

Argument Structure and Analysis AQ is a multidimensional concept assessed through logical, rhetorical, and dialectical aspects (Wachsmuth et al., 2017). Logical quality evaluates coherence via premise believability, relevance, and sufficiency. Rhetorical quality incorporates credibility, emotional appeal, and clarity (Ziegenbein et al., 2023; Ivanova et al., 2024). Dialectical quality measures discourse engagement and argument robustness

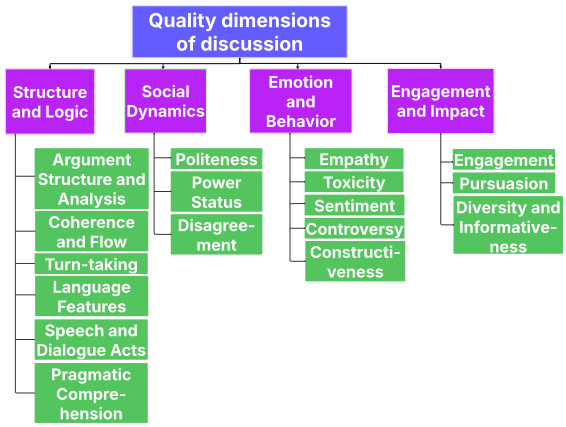

Figure 1: Proposed taxonomy of evaluation aspects.

(Wachsmuth et al., 2017). Empirical studies leverage NLP and ML, and recently LLMs, to automate AQ assessment (Ziems et al., 2024). LLMs demonstrate strong annotation capabilities, performing comparably to human evaluators (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). For a broader review, see the work of Wachsmuth et al. (2024) and Lauscher et al. (2022).

Coherence and Flow Coherence evaluates logical consistency, while flow assesses smooth progression in discussions (Li et al., 2021). Initial efforts in coherence assessment focused on dialogue systems using language features (Zhang et al., 2018a), discourse structure (Barzilay and Lapata, 2008), and graph-based representations (Huang et al., 2020; Zhang et al., 2021). To assess both aspects, LLMs are increasingly used for coherence evaluation at the comment or whole discussion level (Zhang et al., 2024), with proprietary models (e.g., GPT-4) excelling, and fine-tuned open-source models showing promise (Mendonca et al., 2024; Zhang et al., 2023).

Turn-taking Turn-taking patterns (e.g., how often participants speak in a discussion, to which other speakers they tend to reply or not) inform coherence evaluation (Cervone and Riccardi, 2020), constructiveness prediction (Niculae and Danescu-Niculescu-Mizil, 2016), and facilitation analysis (Schroeder et al., 2024). Studies have used entropy (Niculae and Danescu-Niculescu-Mizil, 2016), Gini coefficients (Schroeder et al., 2024), and neural architectures for modeling turn-taking dynamics (Chang and Danescu-Niculescu-Mizil,

2019; Li et al., 2021), as well as turn-taking visualization tools (El-Assady et al., 2017; Hoque and Carenini, 2016).

Language Features Language features have been used to help model content and expression in online discussions (Wilson et al., 1984). Early methods used lexicons for sentiment, toxicity, politeness and collaboration evaluations, aspects closely related to discussion quality (Lawrence et al., 2017; Avalle et al., 2024). Deep learning models leverage word embeddings instead (De Kock and Vlachos, 2021).

Speech and Dialogue Acts Rooted in Speech Act Theory (Austin, 1975; Searle, 1969), dialogue acts characterize dialogue turns (e.g., disagreement, elaboration) to analyze interaction dynamics (Ferschke et al., 2012). Various taxonomies exist, including generic (Stolcke et al., 2000) and domain-specific frameworks (Zhang et al., 2017; Al-Khatib et al., 2018). LLMs, in this case, can serve as dialogue act annotators (Ziems et al., 2024; Cimino et al., 2024; Martinenghi et al., 2024), aiding in the distinction between effective and ineffective discussions based on the patterns of interaction and the communicative strategies and goals of the participants (Ziems et al., 2024).

Pragmatic Comprehension Pragmatic comprehension (how context affects meaning) is essential in human communication, where the intended meaning often differs from the verbalized expression (e.g., in implicature). Humans resolve such ambiguities using social and common-sense knowledge. NLP models need to be evaluated on their pragmatic understanding (Al-Khatib et al., 2018). 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 pragmatic tasks, they struggle in tasks requiring social norm and deep contextual awareness (Hu et al., 2023; Sravanthi et al., 2024).

5.2 Social Dynamics

Politeness Politeness is vital in human interaction and has been studied in relation to conversational derailment (Zhang et al., 2018b) and constructiveness (De Kock and Vlachos, 2021; Zhou et al., 2024). Computational approaches to politeness detection have evolved from ML-based classifiers using domain-independent markers (Danescu-Niculescu-Mizil et al., 2013) to successfully leveraging LLMs for annotation (Zhou et al., 2024; Ziems et al., 2024),

Power and Status Power and status influence conversational dynamics, affecting language use and turn-taking. Low-power individuals tend to mimic high-power speakers' linguistic styles more than vice versa (Danescu-Niculescu-Mizil et al., 2012). Hence, higher status roles can control the flow of discussions, foster social inequalities and hence degrade discussion quality. LLMs perform well in identifying power differentials in discussions (Ziems et al., 2024).

Disagreement Disagreements, when constructive, enhance discussions by fostering deeper understanding (Friess, 2018; De Kock and Vlachos, 2021). Measuring disagreement levels is complex, with frameworks such as Graham's hierarchy (ranging from name calling to refuting the central point; Graham, 2008) and dispute tactics (Walker et al., 2012; Benesch et al., 2016; De Kock et al., 2022) providing structured analyses. LLMs have been successfully employed as dispute tactics annotators (Zhou et al., 2024).

5.3 Emotion and Behavior

Empathy Empathy fosters constructive discussions and is commonly measured as Perceived Empathy (Concannon et al., 2023). Approaches range from coding schemes to linguistic markers (Macagno et al., 2022). Evaluations of LLMs in empathy detection tasks show that empathy understanding is challenging for LLMs (Ziems et al., 2024; Xu and Jiang, 2024).

Toxicity Toxicity in online discussions refers to harmful or disrespectful language that hinders productive discourse (Avalle et al., 2024). Identifying toxicity is vital to maintain constructive communication. LLMs have been proven to be adept at identifying toxicity (Ziems et al., 2024).

Sentiment Sentiment analysis gauges the emotional tone of discussions, which influences the quality of interactions (De Kock and Vlachos, 2021). It helps identify whether discussions are positive, negative, or neutral. GPT-4 has successfully been utilized as a sentiment annotator (Zhou et al., 2024). Ziems et al. (2024) also evaluated LLMs in figurative language identification, including sarcasm, revealing moderate performance and occasional misclassification.

Controversy Controversy arises from divergent viewpoints, leading to polarized exchanges. The spread of political leanings among discussion participants of controversial topics and sentiment distribution analysis are common approaches to measure it (Avalle et al., 2024). Ziems et al. (2024) evaluated GPT-4 in political ideology identification, showing moderate predictive performance.

Constructiveness Constructiveness fosters meaningful dialogue, especially in online discussions, by promoting resolution and cooperation (Shahid et al., 2024). It is signalled by linguistic features (De Kock et al., 2022; Falk et al., 2024) and empirical research develops models to predict it (Zhou et al., 2024). Shahid et al. (2024) found that GPT-4 preferred dialectical over logical arguments in assessing constructiveness. They also found that human participants rated LLM-generated and human-AI co-written comments as significantly more constructive than those written independently by humans.

5.4 Engagement and Impact

Engagement refers to the level of interest and participation in a discussion. It can be assessed by measuring response relevance, balance between questions and statements, interaction flow, and sometimes discussion length (Adomavicius, 2021). Zhang et al. (2024) document that engagement assignment is a challenge task for LLMs.

Persuasion Empirical literature has primarily investigated factors influencing persuasion, including linguistic strategies (e.g., word matching) and interaction dynamics (e.g., back-and-forth engagement (Tan et al., 2016)). Researchers have also examined the semantic types of argumentative components (premises and claims) , such as ethos-based appeals and interpretations(Hidey et al., 2017). Additionally, studies have explored dynamic factors related to topics and discourse (Zeng et al., 2020) and have developed models aimed at accurately predicting persuasion. However, Ziems et al. (2024) highlight the poor persuasiveness annotation capability of LLMs in online argumentative discussions.

Diversity and Informativeness Diversity improves discussion quality by introducing varied perspectives and experiences (Zhang et al., 2024). Informativeness refers to the relevance and value of information shared in a discussion. Zhang et al. (2024) document that LLMs struggle to assess diversity and informativeness.

6 Intervention Strategies

6.1 When to Intervene

Picking the right moment to intervene is a crucial part of effective facilitator 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). It is imperative then, that a facilitator is able to recognize subtle cues hinting towards escalation, in order to defuse the situation (something that even experienced humans 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 event would be a facilitation intervention (Schluger et al., 2022). One way to solve this problem models it as a binary classification task (see 'Conversation Derailment' datasets, §8). Traditional ML models can perform better than baselines on this task, although their performance varies (Falk et al., 2021; Park et al., 2012; Falk et al., 2024; Schluger et al., 2022).

6.2 How to Intervene

There is currently no standard, 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, each of which corresponds to several possible intervention types. These functions roughly correlate with the volunteer moderator roles found in 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).

Facilitators often have to decide what form of coercive measure they will take to make sure the conversation remains healthy, without having to intervene repeatedly. Human interventions typically use an unofficial 'escalation ladder', where the fa-

cilitator will progressively move from standard facilitation tactics to threatening, and finally disciplinary action (Seering, 2020). Disciplinary action should be used as a last resort, since '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). Indeed, it is typically not the first choice of a facilitator (Schluger et al., 2022)

That said, there are also 'softer' kinds of interventions that facilitators frequently use, such as: 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 in structuring their speech (Tsai et al., 2024; Falk et al., 2024).

6.3 Personalized Interventions

Finally, it is worth stressing that intervention strategies should not be applied en masse, without considering the characteristics of each individual. Traditionally, massive application (or threatening) of disciplinary action has led to adverse effects community- and platform-wide (Trujillo and Cresci, 2022; Falk et al., 2021) and 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.

7 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 (such as Kim et al. (2021)), where the model produces automated messages encouraging participation). However, ML-based and rule-based facilitation are not effective enough to meet the high demands of most platforms (Seering, 2020; Schaffner et al., 2024).

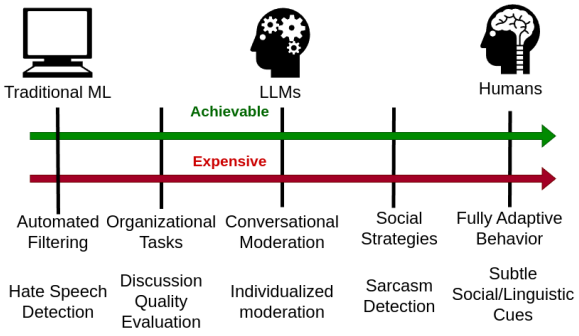

Figure 2: Current capabilities of ML, LLM, and human facilitation. From left to right: new tasks are possible, but the cost of solving them rises proportionally. In-between actors, are tasks that are solved suboptimally by the previous actor; e.g., ML systems can be used for discussion evaluation, but are not particularly effective.

Advances in LLMs enable the development of *facilitation agents* that more actively engage 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 (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 ML models, LLMs, and humans can be found in Figure 2.

7.1 Administrating the discussion

LLMs are well positioned to tackle a wealth of 'administrative' facilitation tasks, which aim to organize a discussion. For instance, facilitators typically periodically summarize the viewpoints of the participants and seek their confirmation in order to both make them feel understood, and present their views to other participants. This iterative summarization is a task which LLMs may be well-equipped to handle (Small et al., 2023; Burton et al., 2024). According to Jin et al. (2024), LLMs bring significant advantages over traditional methods, "notably in the quality and flexibility of the generated texts and the prompting paradigm to alleviate the cost of training deep models". Specifically on discussion summarization, however, Feng and Qin (2022) suggest that the challenges are profound, as discussions contain multiple participants, topic drifts, multiple co-references, diverse interactive signals, and diverse domain terminologies.

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 current Information Retrieval (IR) approaches (Karadzhov et al., 2021).

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

7.2 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 (fake news) (Kang and Qian, 2024; Wang and Chang, 2022). These abilities make LLMs usable as drop-in replacements for traditional ML models for these tasks. They also suggest that conversational LLM facilitation agents may be able to identify, and dynamically adapt to such phenomena properly.

7.3 Fully Automatic LLM-based facilitation

There are indications that LLM chatbots can be used as facilitators in 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., 2024; Martinenghi et al., 2024). Given that relatively simple ML chatbots (Kim et al., 2021), which do not leverage generative text capabilities, succeed at improving discussions, 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), with some applications already showing promise (Cho et al., 2024).

8 Facilitation Datasets

In this section, we provide an overview of the most prominent datasets for online moderation, considering their sizes and their relevance to core facilitation tasks. We propose the following new taxonomy of facilitation datasets: **Conversation Derailment** datasets, where the task is to predict when a conversation escalates and requires facilitator intervention; **User Tactics** datasets, which concern how users position themselves during the discussion; and **Facilitator Interventions** datasets. Some datasets contain information that can be used in multiple tasks. An overview of the surveyed datasets and their categories in our taxonomy can be found in Table 1.

9 LLM Discussion Facilitation Roadmap

Evaluation LLMs can serve as automated discussion quality annotators. Are these annotators infallible? At the moment, the answer is no. There are discussion quality aspects, with specific characteristics, that LLMs cannot annotate in a highly accurate way. Particularly, aspects characterized by highly subjective and complicated expert taxonomies (e.g., empathy), as well as specific AQ dimensions requiring the integration of social norms (e.g., diversity; see §5). However, even human annotations tend to be polarized for such subjective quality dimensions (e.g., toxicity) due to sociodemographic background effects and personal biases (Beck et al., 2024; Sap et al., 2020).

Unlike traditional automated (self-)supervised techniques that require costly hardware and extensive training data, prompted LLMs offer a more affordable alternative for discussion quality annotation without the need for expensive training examples. Yet, as any tool, LLMs should be used with caution. For each quality aspect of the discussion, researchers must choose the proper models (e.g., open/close source, the size, trained or not with reinforcement learning, fine-tuned or not on specialized data). Moreover, prompts specifically engineered for each quality aspect may also be adopted.

Future research should consider practical evaluation frameworks for both the meta-evaluation of discussion quality metrics—assessing their effectiveness, consistency, bias, robustness, and real-world applicability—and for moderation and facilitation interventions. Evaluating such frameworks and facilitation strategies requires rigorous testing in controlled environments. However, real-world evaluations can be costly, time-consuming, and difficult to replicate, making alternative approaches necessary. To this end, synthetic experiments made exclusively with LLM actors may be of great help, although future work will have to prove that synthetic experiments are a reliable proxy for human behavior to a certain degree (Hewitt et al., 2024;

Name	Task		Size	Content
Wikipedia Disputes (De Kock and Vlachos, 2021)	Conversation Derailment	User Tactics	7,425 D	Includes annotations for several 'dispute tactics.'
WikiConv (Hua et al., 2018)	User Tactics		91,000,000 D	Includes moderation meta-data such as comment edits and deletions.
Webis-WikiDebate-18 (Al-Khatib et al., 2018)	User Tactics		6,000,000 D	Graph-based, includes data annotated for argumentation strategies.
Conversations Gone Awry (Zhang et al., 2018b)	Conversation Derailment	User Tactics	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).
DeliData (Karadzhov et al., 2021)	User Tactics	Facilitator Interventions	500 D	Group discussions, includes task-oriented quality measure which may be used to approximate discussion quality.
Wiki-Tactics (De Kock et al., 2022)	Facilitator Interventions		213 D	Based on Wikipedia Disputes, includes moderation action metadata such as comment edits and deletions.
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.

Table 1: Overview of reviewed datasets. Unnamed datasets are referred to by the names of the authors. The size reflects the number of annotated conversations, disregarding unlabeled data. In this table, **D** refers to the number of discussions, while **C**, to the number of individual comments.

Park et al., 2023, 2022, 2024).

Facilitation Intervention types should be adapted to the different legal frameworks and rules of each community/platform. While there are exhaustive surveys on the topic, such as that of Schaffner et al. (2024), there is yet no methodology to train human or synthetic 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, as applied in other NLP tasks (Ulmer et al., 2024; P.Cheng et al., 2024; Park et al., 2022, 2023). 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 facilitation training settings.

10 Conclusions

We surveyed the current literature on online discussion evaluation and facilitation, taking into account two main factors: (a) the need to leverage ideas from the intersection of Social Science and NLP, when it comes to discussion quality evaluation and facilitation strategies, and (b) the current capabilities and potential applications of LLMs. Focusing on thread-like discussions, we proposed a new taxonomy for online discussion quality evaluation. In terms of intervention strategies, effective facilitation is crucial in online discussions to prevent escalation, reduce toxicity, and ensure the conversation remains healthy and productive, with advancements in both human and machine-driven interventions showing significant promise in improving the quality and timeliness of these interventions. 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, opening up new opportunities to foster healthier and more constructive conversations.

11 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. However, since summarization is a vast subfield of NLP, it is only briefly mentioned in this survey.

Moreover, it is important to highlight that most research on facilitation has been conducted solely in English-speaking online spaces. Additionally, 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, in the real world, the majority of online discussions and sometimes deliberations happen in the context of communities, where group dynamics apply. Thus, a complete review of facilitation has to account for the internal dynamics of such communities, as well as the wider role of community facilitators.

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.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.

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. Barzilay and M. Lapata. 2008. Modeling local coherence: An entity-based approach. *Computational Linguistics*, 34(1):1–34.

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.

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.

P. Cheng, T. Hu, H. Xu, Z. Zhang, Y. Dai, L. Han, and N. Du. 2024. 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, I. Roberts, and M. Tomalin. 2023. An interactional account of empathy in human-machine communication. *Human-Machine Communication*, 6.

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. Danescu-Niculescu-Mizil, M. Sudhof, D. Jurafsky, J. Leskovec, and C. Potts. 2013. A computational approach to politeness with application to social factors. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 250–259, Sofia, Bulgaria.

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.

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*.

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.

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.

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.

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

C. Hidey, E. Musi, A. Hwang, S. Muresan, and K. McKeown. 2017. Analyzing the semantic types of claims and premises in an online persuasive forum. In *Proceedings of the 4th Workshop on Argument Mining*, pages 11–21, Copenhagen, Denmark.

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.

L. Huang, Z. Ye, J. Qin, L. Lin, and X. Liang. 2020. GRADE: Automatic graph-enhanced coherence metric for evaluating open-domain dialogue systems. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9230–9240, Online.

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.

R. Ivanova, T. Huber, and C. Niklaus. 2024. Let's discuss! quality dimensions and annotated datasets for computational argument quality assessment. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 20749–20779, Miami, Florida, USA.

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*.

A. Kahane. 2013. Transformative scenario planning: Working together to change the future. *Stanford Social Innovation Review*.

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.

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.

G. Lapesa, E. M. Vecchi, S. Villata, and H. Wachsmuth. 2024. Mining, assessing, and improving arguments in NLP and the social sciences. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024): Tutorial Summaries*, pages 26–32.

A. Lauscher, H. Wachsmuth, I. Gurevych, and G. Glavaš. 2022. Scientia potentia est—on the role of knowledge in computational argumentation. *Transactions of the Association for Computational Linguistics*, 10:1392–1422.

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).

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.

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.

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).

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.

J.S. Park, C.Q. Zou, A. Shaw, B. Mako Hill, C.J. Cai, M.R. Morris, R. Willer, P.L., and M.S. Bernstein. 2024. Generative agent simulations of 1,000 people. *ArXiv*, abs/2411.10109.

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

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. 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 discourse| 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.

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.

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.

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.

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.

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.

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, 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.

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

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.

J. Zeng, J. Li, Y. He, C. Gao, M. Lyu, and Ir King. 2020. What changed your mind: The roles of dynamic topics and discourse in argumentation process. In *Proceedings of The Web Conference 2020*, WWW '20, page 1502–1513, New York, NY, USA. Association for Computing Machinery.

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, Y. Chen, L. F. D'Haro, Y. Zhang, T. Friedrichs, G. Lee, and H. Li. 2021. DynaEval: Unifying turn and dialogue level evaluation. 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 5676–5689, Online.

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.

H. Zhang, Y. Lan, J. Guo, J Xu, and X. Cheng. 2018a. Reinforcing coherence for sequence to sequence model in dialogue generation. In *IJCAI*, pages 4567–4573.

J. Zhang, J. Chang, C. Danescu-Niculescu-Mizil, L. Dixon, Y. Hua, D. Taraborelli, and N. Thain. 2018b. 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.

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

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.

T. Ziegenbein, S. Syed, F. Lange, M. Potthast, and H. Wachsmuth. 2023. Modeling appropriate language in argumentation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4344–4363.

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

DT Deliberative Theory

AQ Argument Quality

B Keywords for Literature Query

C Terminology Background

In this section of the Appendix we provide the relevant literature, as well as our reasoning for choosing and disambiguating certain terms used in this paper (see §2) and which also allow us to define the scope of this survey. The definitions for the terms used in this paper 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 structure 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 for describing 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.

For the sake of conciseness, this survey focuses on real-time moderation and ex-post moderation.

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

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.

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 occupied. 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 obtains another sense when we refer to dialogue systems, a major NLP sub-are, which traditionally includes both task-oriented dialogues but also casual conversation style (Elize-style)[1] "chatbots" (Liu et al., 2023; Sun et al., 2021).

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).

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 dimensions from all the other rhetorical styles. For example, the adversarial dimension of the debate or the argumentative dimension that can be found both in dialogues and deliberation-style conversations.

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., 2018b), as well as their level of engagement (See et al., 2019). It also depends on the type of the discussion, since, as we mentioned in the previous paragraph, some types of rhetoric style, such as deliberative discussions do not always aim to achieve consensus. Given these complexities, we adopt the definition proposed by Raj Prabhu et al. (2021),

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

which views the perceived *discussion quality* as a measure that attempts to quantify interactions by taking into account multiple socio-dimensional aspects of individual experiences and abilities.