Scalable Evaluation of Online Facilitation Strategies via Synthetic Simulation of Discussions

Anonymous ACL submission

Abstract

Limited large-scale evaluations exist for facilitation strategies of online discussions due to significant costs associated with human involvement. An effective solution is synthetic discussion simulations using Large Language Models (LLMs) to create initial pilot experiments. We propose a simple and generalizable LLM-driven methodology to prototype LLM facilitators, and produce high-quality synthetic data without human involvement. We use our methodology to test whether modern facilitation strategies can improve the performance of LLM facilitators. We find that, while LLM facilitators significantly improve synthetic discussions, there is no evidence that the application of modern facilitation strategies leads to further improvements in discussion quality. We also find that small LLMs (such as Mistral Nemo 12B) can perform comparably to larger models (such as LLaMa 70B), and that special instructions must be used for instruction-tuned models to induce toxicity in synthetic discussions. We confirm that each component of our methodology contributes meaningfully to high quality data via an ablation study. We also release an open-source framework XXX[1] (`pip install xxx`), which implements our methodology, and release a large, publicly available dataset containing LLM-generated and LLM-annotated discussions from multiple open-source LLMs.

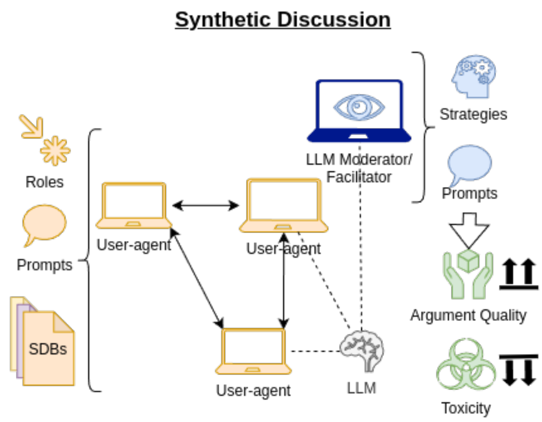


Figure 1: LLM user-agents with distinct SocioDemographic Backgrounds (SDBs) participate in a discussion, while the LLM moderator monitors and attempts to improve the quality of the discussion. We need to design prompts and configurations for both types of LLM agents.

1 Introduction

Research on conversational moderation/facilitation techniques is crucial for adapting to ever-changing and demanding online environments. Relevant work traditionally focused on isolating and removing content (Seering, 2020; Cresci et al., 2022), whereas the current social media environment demands moderation systems to adequately explain their actions and prevent problematic behaviors before they surface (Cho et al., 2024; Seering, 2020; Cresci et al., 2022; Amaury and Stefano, 2022). Facilitation mechanisms are also needed to handle community deliberation and group decision-making (Kim et al., 2021; Seering, 2020). Note that "content moderation" usually involves flagging and removing content, as opposed to "conversational moderation", which is studied in this paper. The terms "facilitation" and "conversational moderation" are otherwise equivalent (Argyle et al., 2023; Korre et al., 2025; Falk et al., 2021) and we treat them as synonyms in this paper.

A major challenge in connecting facilitation research to real-world needs is the substantial costs required both in researching and moderating discussions, due to human participation (Rossi et al., 2024). Many social media platforms overcome this by outsourcing moderation to volunteers or their own users (Matias, 2019; Schaffner et al., 2024), while others support only conventional content moderation using traditional Machine Learning (ML) models, which are not enough in practice

[1] anonymous.4open.science/r/framework-F8E6

(Horta Ribeiro et al., 2023; Schaffner et al., 2024). Large Language Models (LLMs) have been hypothesized to be capable of conversational moderation and facilitation tasks, which often require actively participating in the discussions, instead of passively flagging or removing content (Small et al., 2023; Korre et al., 2025).

While studies exist for simulating user interactions in social media (Park et al., 2022; Mou et al., 2024; Törnberg et al., 2023; Rossetti et al., 2024; Balog et al., 2024), and for using artificial facilitators (Kim et al., 2021; Cho et al., 2024), none so far have combined the two approaches. We posit that synthetic simulations can be a cheap and easy way to develop and test preliminary, in silico experiments with LLM facilitators, initial versions of which may be unstable or unpredictable (Atil et al., 2025; Rossi et al., 2024), before testing them in much more costly experiments with human participants. Our work thus asks the following two questions: (1) Can we produce high-quality synthetic discussions, involving alternative facilitation strategies, by crafting an appropriate environment for simulations? (2) Can we boost the effectiveness of LLM moderators (in synthetic discussions) by using prompts aligned with facilitation strategies proposed in modern Social Science research?

We propose a simple and generalizable approach using LLM-driven synthetic experiments for online moderation research, enabling fast and inexpensive model "debugging" and parameter testing (e.g., LLM moderator prompts, instructions) without human involvement (§3) (Fig. 1). An ablation study (§5.2) demonstrates that each step of our methodology meaningfully contributes to generating high-quality synthetic data. Using this methodology, we examine four LLM moderation strategies (including a novel strategy inspired by Reinforcement Learning (RL)) based on current Social Science facilitation research (§4) and compare them with two baselines (LLM facilitators with simplistic facilitation prompts). LLMs are also used to gauge discussion quality (e.g., argument quality, toxicity).

Our analysis reveals two key findings (§5): (1) the presence of LLM facilitators has a positive and statistically significant influence on the quality of synthetic discussions, and (2) facilitation strategies inspired by Social Science research often do not manage to outperform simpler baselines. Furthermore, we release XXXan open-source Python framework for generating and evaluating synthetic discussions, alongside a large, publicly available dataset comprising automatically evaluated synthetic discussions (§6). We use open-source LLMs and include all relevant configurations in order to make our study as reproducible as possible (see §A.3, §A.4).

2 Background and Related Work

2.1 LLMs as Human Subjects

When conducting social experiments with LLMs instead of human subjects, it is imperative to know how representative results can be. Grossmann et al. (2023) argue that synthetic agents have the potential to eventually replace human participants, a perspective shared by other researchers (Törnberg et al., 2023; Argyle et al., 2023). Indeed, LLMs have demonstrated emergent complex social behaviors (Park et al., 2023; Marzo et al., 2023; Leng and Yuan, 2024; Abdelnabi et al., 2024; Abramski et al., 2023), and are able to infer survey responses from SDBs (Hewitt et al., 2024) and personalized interviews (Park et al., 2024).

However, significant limitations of LLMs remain in the context of Social Science experiments. Issues include dataset contamination; undetectable behavioral hallucinations (Rossi et al., 2024); sociodemographic, statistical and political biases (Anthis et al., 2025; Hewitt et al., 2024; Rossi et al., 2024), often amplified during discussions (Taubenfeld et al., 2024); unreliable survey responses (Jansen et al., 2023; Bisbee et al., 2024; Neumann et al., 2025); inconsistent annotations (Gligori'c et al., 2024); non-deterministic outputs (Atil et al., 2025), especially in closed-source models (Bisbee et al., 2024); and excessive agreeableness due to alignment procedures (Park et al., 2023; Anthis et al., 2025; Rossi et al., 2024). Despite these shortcomings, researchers frequently anthropomorphize LLM agents (Rossi et al., 2024), obscuring the true causes of their behavior (Anthis et al., 2025; Zhou et al., 2024a).

Our study must thus be conservative towards the generalizability of our results to discussions with human participants. We stress that our methodology is designed for "debugging" and exploring artificial facilitators in silico, before testing them in much more costly experiments with human participants. Experiments with real participants, however, are ultimately needed, and we leave them for future work.

2.2 Evaluating Discussion Quality

Synthetic discussions often degrade rapidly without human interaction, exhibiting repetitive, low-quality content (Ulmer et al., 2024). However, research on quantifying synthetic data quality is currently limited.

Balog et al. (2024) introduce metrics utilizing comparisons with human data, but this approach depends on datasets with the same topics, and lacks scientific grounding since believable LLM outputs do not necessarily lead to behavior simulation (Rossi et al., 2024). Their most generalizable metric—a vague "coherence" score—is LLM-annotated without theoretical support. Kim et al. (2021) rely on post-discussion surveys and lexical diversity to estimate the number of diverse opinions.

Alternatively, Ulmer et al. (2024) propose *"Diversity"*, which penalizes repeated sequences between comments in a discussion:

$$div(d) = 1 - \frac{2}{N_d(N_d - 1)} \sum_{i=1}^{N_d} \sum_{j=i+1}^{N_d} R(c(i,d), c(j,d)) \quad (1)$$

where R is the ROUGE-L F1 score[2] (Lin, 2004), and N_d the length (in comments) of discussion d.

Low diversity points to pathological problems (e.g., LLMs repeating previous comments). Extremely high diversity scores, on the other hand, may point to a lack of interaction between participants; a discussion in which participants engage with each other will feature some lexical overlap (e.g., common terms, paraphrasing points of other participants). We can instead compare the distribution of *diversity* scores for synthetic discussions with that measured on sampled human discussions. This allows us to estimate the extent to which synthetic discussions approximate real-world content variety and participant interaction.

Besides metrics for the quality of synthetic data, we also need metrics that can quantify how "well" a discussion is going from a human standpoint. We choose Toxicity for two reasons: Prompting LLMs for toxicity detection is reliable (Kang and Qian, 2024; Wang and Chang, 2022; Anjum and Katarya, 2024), and toxicity can inhibit online and deliberative discussions (De Kock et al., 2022; Xia et al., 2020).[3]

[2] We use the rouge-score package in our analysis.

[3] We note that this is not always true (Avalle et al., 2024).

2.3 Synthetic Discussions

Synthetic discussion systems include synthetic clones of Reddit (Park et al., 2022), Twitter/X (Mou et al., 2024), social media in general (Törnberg et al., 2023; Rossetti et al., 2024) as well as games (Park et al., 2023) and social experiments (Zhou et al., 2024b).

Balog et al. (2024) introduce their own methodology to produce synthetic discussions, where they extract topics and comments from real-world online discussions, and prompt an LLM to continue them. Unlike our approach, they do not use LLM user-agents to model conversational dynamics, nor do they model the presence of facilitators. Their methodology faces challenges when LLMs generate malformed metadata, for which they offer no solution, and relies on the existence of suitable human discussion datasets.

Ulmer et al. (2024) create synthetic discussions between two participants; an agent (who controls the environment) and a client (who interacts with the agent). They then filter the generated discussions and use them as training data to further fine-tune the agent LLM for a specific task. Their approach however does not model the existence of multiple clients (users), nor is it applied on online discussion facilitation. Our proposed methodology can be modelled as a generalization of their paradigm; an agent (moderator) converses with multiple clients (non-moderator users).

Finally, Abdelnabi et al. (2024) create synthetic negotiations with multiple agents having various agendas and responsibilities. Our work can be modelled as a domain shift of their methodology from negotiations, to discussion facilitation; participants with different motivations (i.e., normal users, trolls, long-standing community members), interact with themselves and a stakeholder holding veto power (facilitator) who presides over the discussion.

2.4 LLM Facilitation

Unlike traditional ML models, LLMs can actively facilitate discussions (Korre et al., 2025). They can warn users for rule violations (Kumar et al., 2024), monitor engagement (Schroeder et al., 2024), aggregate diverse opinions (Small et al., 2023), provide translations and writing tips, which is especially useful for marginalized groups (Tsai et al., 2024). These capabilities suggest that LLMs may be able to assist or even replace human facilitators in many tasks (Seering, 2020).

Moderator chatbots have shown promise; Kim et al. (2021) demonstrated that simple rule-based models can enhance discussions, although their approach was largely confined to organizing the discussion based on the"think-pair-share" framework (Nik Ahmad, 2010; Navajas et al., 2018), and balancing user activity. Cho et al. (2024) use LLM facilitators in human discussions, with moderation strategies based on Cognitive Behavioral Therapy and the work of Rosenberg and Chopra (2015). They show that LLM facilitators can provide "specific and fair feedback" to users, although they struggle to make users more respectful and cooperative. In contrast to both works, our work uses exclusively LLM participants (and LLM facilitators), and tests them in an explicitly toxic and challenging environment.

3 Methodology

3.1 Defining Synthetic Discussions

We assume that the h most recent preceding comments at any given point in the discussion provide sufficient context for the LLM agents (users, facilitators, annotators) (Pavlopoulos et al., 2020). This approach eliminates the need for additional mechanisms such as summarization (Balog et al., 2024), LLM self-critique (Yu et al., 2024), or memory modules (Vezhnevets et al., 2023), resulting in reduced computational overhead and a more transparent, explainable system.

Additionally, we assume that three key functions define the structure of synthetic discussions:

- Underlying model ($LLM(\cdot)$).
- Turn-taking function (t): Determines which user speaks at each turn.
- Prompting function (ϕ): Provides each participant with a personalized instruction prompt, including information such as name and SDB.

We can then model a synthetic comment c at position i of a discussion d recursively as:

$$c(d,i) = LLM(\phi(t(d,i)) \mathbin{+\!\!+} [c(d,j)]_{i-h}^{i-1}) \quad (2)$$

where $\mathbin{+\!\!+}$ is the string concatenation operator, h is the context length of the LLM user-agent (how many preceding comments they can "remember"), and $[c(d,j)]_{i-h}^{i-1} \ldots]$ denotes the concatenation of the previous h comments.

Our formulation of synthetic discussions not only keeps the system simple, but also enables controlled experimentation with various alternatives for each of the three functions (Section 5.2).

3.2 Turn Taking

In online discussions, users do not take turns uniformly, nor do they randomly select which comments to respond to. Instead, they often create "comment chains" where they follow up on responses to their own previous comments. To simulate this, our proposed function chooses between the preceding user and another random user for each turn in the discussion:

$$t(i) = \begin{cases} \mathit{unif}(U) & i = 1, i = 2 \\ \mathit{unif}(U / \{t(i-1)\}) & i > 2, p = 0.6 \\ t(i-2) & i > 2, p = 0.4 \end{cases} \quad (3)$$

where U is the set of all non-facilitator users, *unif* is a function sampling from the uniform distribution, and p represents the probability of the corresponding option being selected. When a facilitator is present, t alternates between picking a normal user and the facilitator (the latter decides whether to respond to or not—the LLM producing an empty string is equivalent to not responding).

3.3 Prompting

SocioDemographic Backgrounds (SDBs) have proven promising in generating varied responses, and alleviating the Western bias exhibited by LLMs (Burton et al., 2024). We generate characteristics for 30 LLM user personas with unique SDBs by prompting a GPT-4 model (OpenAI et al., 2024) (§A.4.1). We do not explicitly include political positions in the prompts of the participants, since instruction-tuned LLMs have been shown to be inherently left-leaning—which can not be alleviated by prompting alone (Taubenfeld et al., 2024)—and research in the field has predominantly occupied Western politics (Taubenfeld et al., 2024; Potter et al., 2024; Rozado, 2024; Pit et al., 2024). Following the paradigm presented by Abdelnabi et al. (2024), we assign roles to non-facilitator user-agents, which inform their incentives for participating in the discussion (e.g., helping the community or disrupting discussions). Each role was mapped to specific instructions (§A.4.3). We create three roles for users: neutral, trolls, and community-focused users. Finally, we select a user instruction prompt (§A.4.2) which instructs participants that repeatedly toxic posts *should* influence their behavior.

4 Experimental Setup

4.1 Moderation Strategies

We test four different facilitation strategies,[4] along with two naive ones that serve as baselines for discussion facilitation:

1. **No Moderator**: A *baseline* where no facilitator is present.
2. **No Instructions**: A *baseline* where a LLM facilitator is present, but is provided only with basic instructions (e.g., "You are a moderator, keep the discussion civil").
3. **Moderation Game**: Our proposed *experimental* strategy, inspired by Abdelnabi et al. (2024) (§2.3). Instructions are formulated as a game, where the facilitator tries to maximize their scores by arriving at specific outcomes (e.g., "User is toxic: -5 points, User corrects behavior: $+10$ points"). No actual score is being kept; they exist to act as indications for how desirable an action or outcome is. The other participants are not provided with scores, nor are they aware of the game rules.
4. **Rules Only**: A *real-life* strategy where the prompt is adapted from LLM alignment guidelines (Huang et al., 2024). This provides the facilitator with a set of rules to uphold, without specifying how to uphold them (e.g, "Be fair and impartial, assist users, don't spread misinformation").
5. **Moderation guidelines**: A *real-life* strategy based on guidelines given to human facilitators of Cornell e-Rulemaking Initiative (CeRI) (eRulemaking Initiative, 2017). For example, "Stick to a maximum of two questions, use simple and clear language, deal with off-topic comments").
6. **Facilitation guidelines**: A *real-life* strategy based on the human facilitation guidelines used by the MIT Center for Constructive Communications (White et al., 2024). For example, "Do not make decisions, be a guide, provide explanations").

4.2 Technical Details

For toxicity annotation, we use ten LLM annotator-agents controlled by a model already used in prior work (LLaMa3.1 70B) (Kang and Qian, 2024). Each annotator's prompt includes SDBs distinct from the ones provided to the users, annotation instructions, and few-shot examples (§A.3). Each annotator is tasked with annotating all comments in each discussion once.

We use three open-source models (in Eq 2) from different families and of different sizes: LLaMa 3.2 (70B), Qwen2.5 (33B) and Mistral Nemo (12B). We select the instruction-tuned variants and quantize them to 4 bits, due to our limited resources. The original and ablation experiments were collectively completed within roughly four weeks of computational time, using two Quadro RTX 6000 GPUs. The execution script is available in the project's repository[5]. The automated discussion generation is detailed in §A.2.

5 Results

5.1 Main findings

LLM facilitators significantly improve synthetic discussions. As is shown in Fig. 4, comments in unmoderated discussions exhibit significantly worse toxicity (ANOVA $p < .000$).[6]

The effect of LLM facilitators is amplified over time under all strategies when compared to unmoderated discussions. Table 1 demonstrates that, for example, with *Mod. Guid.*, conversations begin with 0.277 lower toxicity, and the *Mod. Guid.* $\times$ *time* interaction shows that with each additional dialogue turn, toxicity decreases on average by another 0.023 points.

Sophisticated facilitation strategies do not qualitatively further improve synthetic discussions. The impact of the "Rules Only", "Moderation" and "Facilitation Guidelines" strategies (§4.1) is marginal, and sometimes even not statistically significant compared to the second baseline ("No Instructions") (Fig. 4). This suggests that out-of-the-box LLMs may be unable to effectively use which would enable them to effectively use these advanced instructions, verifying recent research demonstrating important limitations in LLM facilitators (Cho et al., 2024).

LLM facilitators choose to intervene far too frequently. Fig. 2 demonstrates that LLM facilitators intervene at almost any opportunity, even though they are instructed to only do so when necessary. Additionally, a qualitative look through the

[4]The exact prompts used per strategy are in §A.4.4.

[5]anonymous.4open.science/r/experiments-B27D

[6]The large size of our dataset allows the use of parametric tests.

dataset reveals that LLM user-agents exhibit atypical tolerance for excessive moderator interventions. Humans in contrast, typically become irritated and more toxic after repeated, unneeded interventions (Schaffner et al., 2024; Amaury and Stefano, 2022; Schluger et al., 2022; Cresci et al., 2022).

Mistral and Qwen generate discussions more aligned with human diversity scores, despite being significantly smaller than the LLaMa model. As is shown in Fig. 5a, Qwen demonstrated the highest diversity among the evaluated models, indicating limited participant interaction (§2.2), followed by Mistral Nemo and LLaMa. However, none of the models closely matched the diversity observed in human discussions. LLaMa's lower diversity validates prior research suggesting that highly aligned LLMs struggle to replicate human dynamics (Park et al., 2023; Leng and Yuan, 2024). Alternatively, it can also be attributed to its longer average comment length (Fig. 6a); we find that there is a statistically significant negative correlation between comment length and diversity in synthetic discussions ($p < .000$), although we can not verify this pattern in human-generated texts ($p = 0.775$). Despite these differences, the performance gaps between models are relatively small; notably, smaller models like Mistral generate synthetic data of comparable quality to that produced by larger models such as Qwen and LLaMa.

Specialized instruction prompts are essential for eliciting toxic behavior in instruction-tuned LLMs. Our instruction prompt for the participants (§3.3) incentivizes them to react to toxic behavior. Indeed, discussions involving "Troll" user-agents, led to increased toxicity among *other* participants, even when moderated under the "No Instructions" strategy (Fig. 3, Student's t-test, $p < .000$). This effect diminishes when we remove these instructions (Fig. 3).

5.2 Ablation Study

In order to assess the impact of each component of our proposed methodology, we generated eight synthetic discussions per ablation experiment, using a single model, Qwen, to limit computational cost. We evaluated the diversity of these generated ablated discussions by comparing their diversity scores (cf. §2.2) with i) discussions in our original dataset produced solely by the Qwen model; and ii) human discussions from the CeRI "Regulation

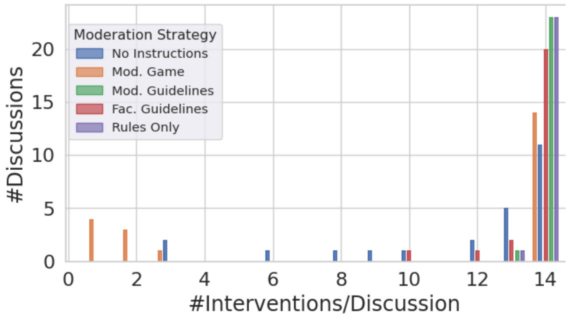


Figure 2: Histogram of interventions by LLM moderators. The maximum number of interventions is 14.

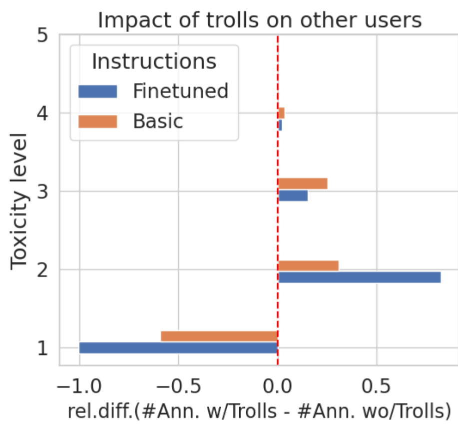


Figure 3: Relative differences in number of annotations per Toxicity of synthetic discussions, when comments by troll users are excluded. We compare between our specialized and a basic instruction prompt.

Room" dataset[7], which includes moderated online deliberative discussions for ten diverse topics.

5.2.1 Effects of Turn Taking Functions

Our proposed turn-taking function meaningfully improves the quality of synthetic data. We compare our turn-taking function (§3.2) to two baselines: Round Robin (participants speaking one after the other, then repeating) and Random Selection (uniformly sampling one of the participants each time). Fig. 5b demonstrates that no single function fully approximates human diversity scores (all distributions diverge from the blue—human—distribution). However, unlike our own function, both baselines feature extremely high diversity. Additionally, Fig. 6b demonstrates that comments in discussions following our turn-taking function closely follow the length of human

[7] `http://archive.regulationroom.org` Disclaimer: Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the CeRI

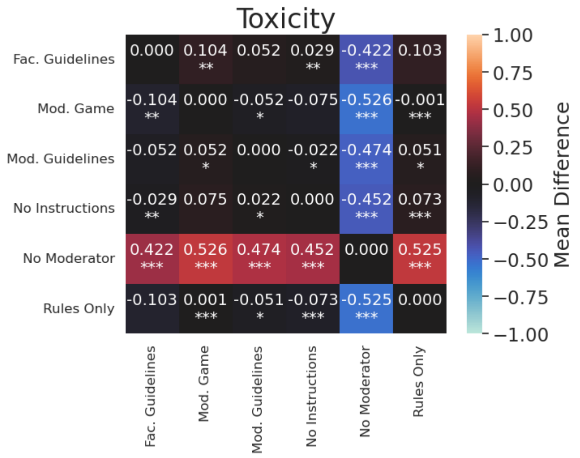


Figure 4: Mean difference of Toxicity between pairs of facilitation strategies. When the value of a cell at row i and column j is x, strategy i leads to overall worse (negative values) or better (positive values) toxicity compared to j for an average of x points in a scale of $1-5$. For each comparison, we use a pairwise Student t-test; p-values are shown as asterisks ($\cdot p < 0.1$, $^{*}\ p < 0.05$, $^{**}\ p < 0.01$, $^{***}\ p < 0.001$).

Variable	Toxicity
Intercept	2.164^{***}
Fac. Guid.	-0.230^{***}
Mod. Guid.	-0.277^{***}
RL Game	-0.435^{***}
No Instructions	-0.426^{***}
Rules Only	-0.461^{***}
time	-0.012^{**}
Fac. Guid×time	-0.023^{***}
Mod. Guid×time	-0.023^{***}
RL Game×time	-0.011^{*}
No Instructions×time	-0.003
Rules Only×time	-0.008

$\cdot p < 0.1$, $^{*}\ p < 0.05$, $^{**}\ p < 0.01$, $^{***}\ p < 0.001$

Table 1: Ordinary Least Squares (OLS) regression coefficients for Toxicity ($Adj.R^2 = 0.054$). *"Time"* denotes dialogue turn, reference factor is *"No moderator"*.

discussions.

5.2.2 Effects of User Prompting

We conduct three separate experiments in which user-agents (excluding moderators) are subjected to one of the following conditions at a time: (1) no assigned SDBs, (2) no assigned roles, or (3) only a basic instruction prompt given (§A.4.2).

SDBs, roles and our instruction prompt increase the quality of synthetic data. Fig. 5c illustrates that although our proposed methodology—incorporating SDBs, roles, and specialized instruction prompts—does not achieve discussions with diversity scores comparable to human ones, replacing any of the above results in a notable deterioration. For instance, omitting SDBs (denoted as "No SDBs" and represented by the red distribution in Fig. 5c) causes the majority of discussions to exhibit maximum diversity—one—indicating a significant loss in participant interaction. This decline is analogous to the effects observed when modifying the turn-taking function. Also similarly to the turn-taking ablation study, our proposed methodology w.r.t. prompts, features comments that best emulate observed human comment length (Fig. 6c).

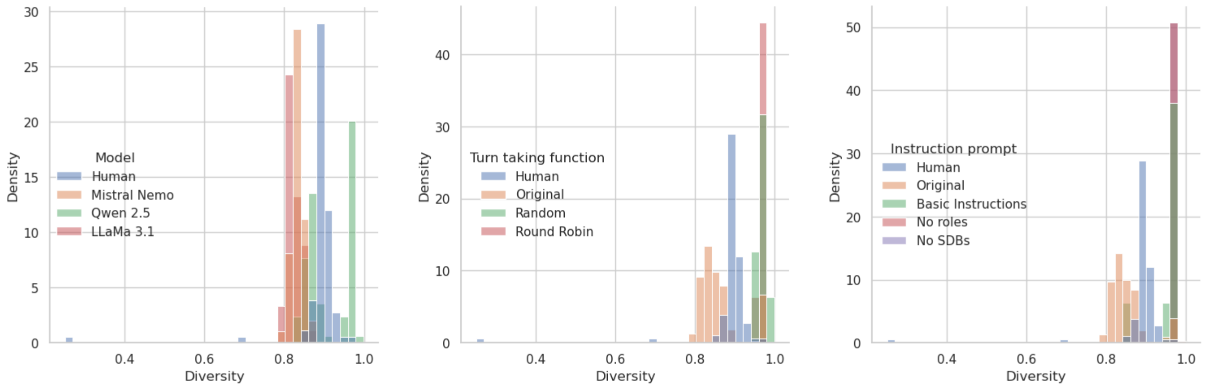
6 Datasets & Software

We introduce XXX[8] an open-source, lightweight, purpose-built framework for managing, annotating, and generating synthetic discussions. Key features include:

- Three core functions: generating, running, and annotating randomized discussion experiments according to provided parameters.
- Built-in fault tolerance (automated recovery and intermittent saving) and file logging to support extended experiments.
- Easy installation via PIP (`pip install xxx`).

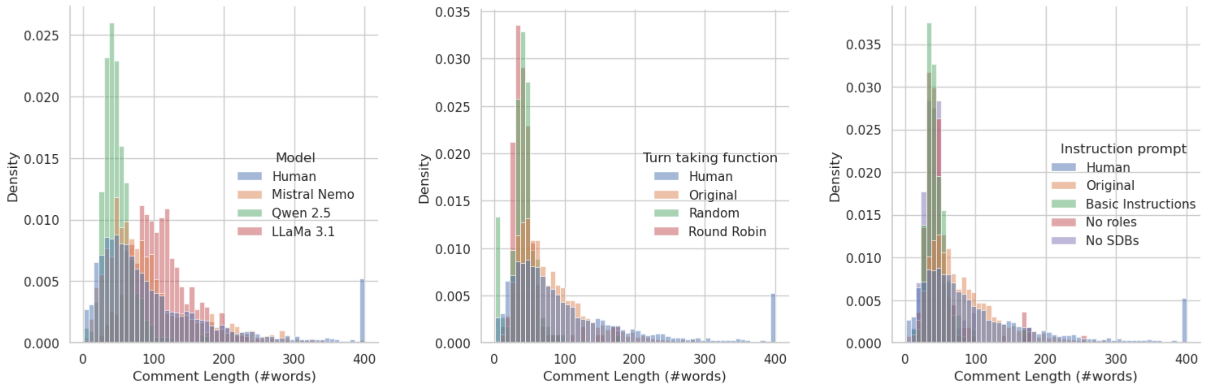
We also release a dataset of synthetic discussions annotated by LLMs for toxicity and argument quality. It can serve as a valuable resource for benchmarking how LLM facilitators would behave according to different facilitation strategies, as well as for further finetuning LLMs, as generally showcased by Ulmer et al. (2024). The supplementary ablation dataset, as well as the code for the analysis and the graphs present in this paper, can be found in the project repository[9]. **Warning: The datasets by their nature contain offensive and hateful speech.**

[8] anonymous.4open.science/r/framework-F8E6
[9] anonymous.4open.science/r/experiments-B27D



(a) Model (b) Turn-taking function t (c) Promoting function ϕ

Figure 5: Diversity (§2.2) distribution for each discussion by LLM (§4.2), turn-taking function t (§3.2), and prompting function ϕ used (§3.3).



(a) Model (b) Turn-taking function t (c) Prompting function ϕ

Figure 6: Comment length for each discussion by LLM (§4.2), turn-taking function t (§3.2), and prompting function ϕ used (§3.3). For ease of comparison, comments above 400 words are marked at the end of the x-axis.

7 Conclusions and Future Work

Our study is the first to apply synthetic data generation to the field of online discussion facilitation. We proposed a simple and generalizable methodology that enables researchers to inexpensively conduct pilot facilitation experiments using exclusively synthetic LLMs. We also conducted an ablation study to demonstrate that each component of our methodology contributes to the production of higher-quality synthetic data.

We created an open-source Python Framework, called XXX, that applies this methodology to hundreds of experiments, which we used to create and publish a large-scale synthetic dataset. Using this dataset, we compared the effectiveness of six moderation strategies and baselines for LLM moderators, elicited from current facilitation research.

Using XXX, we demonstrated that (1) LLM moderators significantly improve the quality of synthetic discussions; (2) established human moderation/facilitation guidelines often do not surpass simple baselines with regard to toxicity and Argument Quality (AQ); (3) smaller LLMs such as Mistral Nemo (12B) can be sufficient for generating high-quality synthetic data; (4) specialized instruction prompts may be needed for instruction-tuned models to feature toxic comments in synthetic discussions.

Future work should identify additional robust quality metrics to evaluate the utility of synthetic data, and examine the applicability of findings obtained on synthetic data (e.g., regarding optimal facilitation strategies) to discussions involving humans. It would also be interesting to explore whether non-instruction-tuned models can generate synthetic discussions that are more aligned with observed human behaviors (Anthis et al., 2025). Finally, synthetic discussion simulations may have the potential to train human moderators before exposing them to real-world discussions.

8 Limitations

Due to limited research in the area, our analysis only uses one synthetic discussion quality metric to gauge data quality. Additionally, while we investigate the impact of facilitation strategies in synthetic discussions, we cannot claim that the behavior of LLM users and facilitator-agents is representative of human behavior. This claim can be scarcely made in Social Science studies involving LLM subjects (Rossi et al., 2024; Zhou et al., 2024a)—as discussed in §2.1.

Furthermore, our experimental setup makes several assumptions that may affect the generalizability of our findings. We examine only three LLMs, assume a maximum of one facilitator per discussion, and use a turn-taking algorithm that overlooks contextual factors like relevance and emotional engagement, which are crucial in human interactions. Moreover, we do not account for the fact that humans may behave differently when knowing they are interacting with LLMs instead of humans. Our methodology also does not take into account interactions where the user-agents and moderator-agents are based on different LLMs (cf. Eq 2). Finally, our analysis partly relies on LLM-generated annotations, potentially introducing known biases associated with LLM annotation (§A.3).

9 Ethical Considerations

Synthetic discussions involving LLMs could be exploited by malicious actors to make LLM user-agents more capable at performing unethical tasks (Majumdar et al., 2024; Marulli et al., 2024). Such actors could adapt our methodology to maximize toxicity, disrupt human discussions, or learn to circumvent moderation mechanisms to propagate misinformation or spread specific agendas. Notably, LLMs currently lack robust defenses against these types of attacks (Li et al., 2025), although ongoing research is addressing these vulnerabilities (Wang et al., 2025).

Even in non-malicious contexts, researchers deploying LLM moderators in real-world communities must do so with transparency and explicit community consent. The undisclosed use of LLM agents can erode trust, be perceived as manipulative (Retraction-Watch, 2025), and potentially violate regulatory standards such as the EU AI Act (European Parliament and Council, 2024). Furthermore, the inherent biases within LLMs risk skewing moderation systems towards the predominant demographics best represented in their training data, often at the expense of disadvantaged or underrepresented groups (Rossi et al., 2024; Anthis et al., 2025; Burton et al., 2024). While the use of SDB prompts is a necessary step toward inclusivity, it remains insufficient for verifiably equitable representation (Rossi et al., 2024).

Additionally, our methodology is designed around batch production of synthetic discussions, each of which necessitates multiple LLM inference calls. While significantly more affordable and environmentally friendly than experiments involving humans (given the carbon footprint associated with humans (Ren et al., 2024)), the potential of our methodology to scale experiments by orders of magnitude may still have non-trivial, adverse environmental effects (Ding and Shi, 2024).

Finally, it is crucial to acknowledge that while LLMs can approximate aspects of human behavior, they do not reliably replicate it (§2.1). Consequently, this research should be viewed as a foundation for pilot experiments, and conclusions about human behavior should be drawn with caution when based solely on synthetic data.

References

Sahar Abdelnabi, Amr Gomaa, Sarath Sivaprasad, Lea Schönherr, and Mario Fritz. 2024. Cooperation, competition, and maliciousness: Llm-stakeholders interactive negotiation. *Preprint*, arXiv:2309.17234.

Katherine Abramski, Salvatore Citraro, Luigi Lombardi, Giulio Rossetti, and Massimo Stella. 2023. Cognitive network science reveals bias in gpt-3, gpt-3.5 turbo, and gpt-4 mirroring math anxiety in high-school students. *Big Data and Cognitive Computing*, 7(3).

T. Amaury and C. Stefano. 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.

Anjum and Rahul Katarya. 2024. Hate speech, toxicity detection in online social media: a recent survey of state of the art and opportunities. *International Journal of Information Security*, 23(1):577–608.

Jacy Reese Anthis, Ryan Liu, Sean M. Richardson, Austin C. Kozlowski, Bernard Koch, James Evans, Erik Brynjolfsson, and Michael Bernstein. 2025. Llm social simulations are a promising research method. *Preprint*, arXiv:2504.02234.

Lisa P Argyle, Christopher A Bail, Ethan C Busby, Joshua R Gubler, Thomas Howe, Christopher Rytting, Taylor Sorensen, and David 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.

Berk Atil, Sarp Aykent, Alexa Chittams, Lisheng Fu, Rebecca J. Passonneau, Evan Radcliffe, Guru Rajan Rajagopal, Adam Sloan, Tomasz Tudrej, Ferhan Ture, Zhe Wu, Lixinyu Xu, and Breck Baldwin. 2025. Non-determinism of "deterministic" llm settings. *Preprint*, arXiv:2408.04667.

Michele Avalle, Niccolò Di Marco, Gabriele Etta, Emanuele Sangiorgio, Shayan Alipour, Anita Bonetti, Lorenzo Alvisi, Antonio Scala, Andrea Baronchelli, Matteo Cinelli, and Walter Quattrociocchi. 2024. Persistent interaction patterns across social media platforms and over time. *Nature*, 628:582 – 589.

Krisztian Balog, John Palowitch, Barbara Ikica, Filip Radlinski, Hamidreza Alvari, and Mehdi Manshadi. 2024. Towards realistic synthetic user-generated content: A scaffolding approach to generating online discussions. *Preprint*, arXiv:2408.08379.

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

J. W. Burton, E. Lopez-Lopez, S. Hechtlinger, and 1 others. 2024. How large language models can reshape collective intelligence. *Nature Human Behaviour*, 8:1643–1655.

Jonathan P. Chang and Cristian Danescu. 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.

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.

Stefano Cresci, Amaury Trujillo, and Tiziano Fagni. 2022. Personalized interventions for online moderation. In *Proceedings of the 33rd ACM Conference on Hypertext and Social Media*, HT '22, page 248–251, New York, NY, USA. Association for Computing Machinery.

Christine De Kock, Tom Stafford, and Andreas 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. Association for Computational Linguistics.

Yi Ding and Tianyao Shi. 2024. Sustainable llm serving: Environmental implications, challenges, and opportunities : Invited paper. In *2024 IEEE 15th International Green and Sustainable Computing Conference (IGSC)*, pages 37–38.

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

European Parliament and Council. 2024. Regulation (eu) 2024/1689 of the european parliament and of the council of 13 june 2024 laying down harmonised rules on artificial intelligence and amending certain union legislative acts (artificial intelligence act). https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32024R1689. OJ L 2024/1689, 12.7.2024.

Neele Falk, Iman Jundi, Eva Maria Vecchi, and Gabriella 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. Association for Computational Linguistics.

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

Kristina Gligori'c, Tijana Zrnic, Cinoo Lee, Emmanuel J. Candes, and Dan Jurafsky. 2024. Can unconfident llm annotations be used for confident conclusions? *ArXiv*, abs/2408.15204.

Igor Grossmann, Matthew Feinberg, Dawn Parker, Nicholas Christakis, Philip Tetlock, and William Cunningham. 2023. Ai and the transformation of social science research. *Science (New York, N.Y.)*, 380:1108–1109.

Ivan Habernal and Iryna Gurevych. 2016. Which argument is more convincing? analyzing and predicting convincingness of web arguments using bidirectional LSTM. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1589–1599, Berlin, Germany. Association for Computational Linguistics.

Luke Hewitt, Ashwini Ashokkumar, Isaias Ghezae, and Robb Willer. 2024. Predicting results of social science experiments using large language models. Equal contribution, order randomized.

Manoel Horta Ribeiro, Justin Cheng, and Robert West. 2023. Automated content moderation increases adherence to community guidelines. In *Proceedings of the ACM Web Conference 2023*, WWW '23, page 2666–2676, New York, NY, USA. Association for Computing Machinery.

Saffron Huang, Divya Siddarth, Liane Lovitt, Thomas I. Liao, Esin Durmus, Alex Tamkin, and Deep Ganguli. 2024. Collective constitutional ai: Aligning a language model with public input. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, page 1395–1417, New York, NY, USA. Association for Computing Machinery.

Bernard J. Jansen, Soon gyo Jung, and Joni Salminen. 2023. Employing large language models in survey research. *Natural Language Processing Journal*, 4:100020.

Hankun Kang and Tieyun 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. Association for Computational Linguistics.

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).

Hyukhun Koh, Dohyung Kim, Minwoo Lee, and Kyomin Jung. 2024. Can LLMs recognize toxicity? a structured investigation framework and toxicity metric. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 6092–6114, Miami, Florida, USA. Association for Computational Linguistics.

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

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.

Yan Leng and Yuan Yuan. 2024. Do llm agents exhibit social behavior? *Preprint*, arXiv:2312.15198.

Ang Li, Yin Zhou, Vethavikashini Chithrra Raghuram, Tom Goldstein, and Micah Goldblum. 2025. Commercial llm agents are already vulnerable to simple yet dangerous attacks. *Preprint*, arXiv:2502.08586.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Durjoy Majumdar, Arjun S, Pranavi Boyina, Sri Sai Priya Rayidi, Yerra Rahul Sai, and Suryakanth V Gangashetty. 2024. Beyond text: Nefarious actors harnessing llms for strategic advantage. In *2024 International Conference on Intelligent Systems for Cybersecurity (ISCS)*, pages 1–7.

Fiammetta Marulli, Pierluigi Paganini, and Fabio Lancellotti. 2024. The three sides of the moon llms in cybersecurity: Guardians, enablers and targets. *Procedia Computer Science*, 246:5340–5348. 28th International Conference on Knowledge Based and Intelligent information and Engineering Systems (KES 2024).

Giordano De Marzo, Luciano Pietronero, and David Garcia. 2023. Emergence of scale-free networks in social interactions among large language models. *Preprint*, arXiv:2312.06619.

Jorge Nathan Matias. 2019. The civic labor of volunteer moderators online. *Social Media + Society*, 5.

Xinyi Mou, Zhongyu Wei, and Xuanjing Huang. 2024. Unveiling the truth and facilitating change: Towards agent-based large-scale social movement simulation. *Preprint*, arXiv:2402.16333.

J. Navajas, T. Niella, and G. et al. Garbulsky. 2018. Aggregated knowledge from a small number of debates outperforms the wisdom of large crowds. *Nature Human Behaviour*, 2:126–132.

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

Nik Azlina Nik Ahmad. 2010. Cetls : Supporting collaborative activities among students and teachers through the use of think- pair-share techniques. *International Journal of Computer Science Issues*, 7.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*.

Joon Sung Park, Lindsay Popowski, Carrie Cai, Meredith Ringel Morris, Percy Liang, and Michael 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. Association for Computing Machinery.

Joon Sung Park, Carolyn Q. Zou, Aaron Shaw, Benjamin Mako Hill, Carrie Cai, Meredith Ringel Morris, Robb Willer, Percy Liang, and Michael S. Bernstein. 2024. Generative agent simulations of 1,000 people. *Preprint*, arXiv:2411.10109.

John Pavlopoulos and Aristidis Likas. 2024. Polarized opinion detection improves the detection of toxic language. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1946–1958, St. Julian's, Malta. Association for Computational Linguistics.

John Pavlopoulos, Jeffrey Sorensen, Lucas Dixon, Nithum Thain, and Ion Androutsopoulos. 2020. Toxicity detection: Does context really matter? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4296–4305, Online. Association for Computational Linguistics.

Isaac Persing and Vincent Ng. 2015. Modeling argument strength in student essays. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 543–552, Beijing, China. Association for Computational Linguistics.

Pagnarasmey Pit, Xingjun Ma, Mike Conway, Qingyu Chen, James Bailey, Henry Pit, Putrasmey Keo, Watey Diep, and Yu-Gang Jiang. 2024. Whose side are you on? investigating the political stance of large language models. *Preprint*, arXiv:2403.13840.

Yujin Potter, Shiyang Lai, Junsol Kim, James Evans, and Dawn Song. 2024. Hidden persuaders: LLMs' political leaning and their influence on voters. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4244–4275, Miami, Florida, USA. Association for Computational Linguistics.

Shuhan Ren, Bill Tomlinson, Rebecca W. Black, and 1 others. 2024. Reconciling the contrasting narratives on the environmental impact of large language models. *Scientific Reports*, 14:26310.

Retraction-Watch. 2025. Experiment using ai-generated posts on reddit draws fire for ethics concerns. `https://retractionwatch.com/2025/04/28/experiment-using-ai-generated-posts-on-reddit-draws-fire-for-ethics-concerns/`. Accessed: 2025-04-29.

Marshall B Rosenberg and Deepak Chopra. 2015. *Nonviolent communication: A language of life: Life-changing tools for healthy relationships*. PuddleDancer Press.

Giulio Rossetti, Massimo Stella, Rémy Cazabet, Katherine Abramski, Erica Cau, Salvatore Citraro, Andrea Failla, Riccardo Improta, Virginia Morini, and Valentina Pansanella. 2024. Y social: an llm-powered social media digital twin. *Preprint*, arXiv:2408.00818.

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

David Rozado. 2024. The political preferences of llms. *PLOS ONE*, 19(7):1–15.

Brennan Schaffner, Arjun Nitin Bhagoji, Siyuan Cheng, Jacqueline Mei, Jay L Shen, Grace Wang, Marshini Chetty, Nick Feamster, Genevieve Lakier, and Chenhao 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. Association for Computing Machinery.

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. 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).

Christopher T. Small, Ivan Vendrov, Esin Durmus, Hadjar Homaei, Elizabeth Barry, Julien Cornebise, Ted Suzman, Deep Ganguli, and Colin Megill. 2023. Opportunities and risks of llms for scalable deliberation with polis. *ArXiv*, abs/2306.11932.

Amir Taubenfeld, Yaniv Dover, Roi Reichart, and Ariel Goldstein. 2024. Systematic biases in llm simulations of debates. *ArXiv*, abs/2402.04049.

Lily L. Tsai, Alex Pentland, Alia Braley, Nuole Chen, José Ramón Enríquez, and Anka Reuel. 2024. Generative AI for Pro-Democracy Platforms. *An MIT Exploration of Generative AI*. Https://mit-genai.pubpub.org/pub/mn45hexw.

Petter Törnberg, Diliara Valeeva, Justus Uitermark, and Christopher Bail. 2023. Simulating social media using large language models to evaluate alternative news feed algorithms. *Preprint*, arXiv:2310.05984.

Dennis Ulmer, Elman Mansimov, Kaixiang Lin, Justin Sun, Xibin Gao, and Yi Zhang. 2024. Bootstrapping llm-based task-oriented dialogue agents via self-talk. *ArXiv*, abs/2401.05033.

Alexander Sasha Vezhnevets, John P. Agapiou, Avia Aharon, Ron Ziv, Jayd Matyas, Edgar A. Du'enez-Guzm'an, William A. Cunningham, Simon Osindero, Danny Karmon, and Joel Z. Leibo. 2023. Generative agent-based modeling with actions grounded in physical, social, or digital space using concordia. *ArXiv*, abs/2312.03664.

Henning Wachsmuth, Nona Naderi, Yufang Hou, Yonatan Bilu, Vinodkumar Prabhakaran, Tim Alberdingk Thijm, Graeme Hirst, and Benno 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, Valencia, Spain. Association for Computational Linguistics.

Huandong Wang, Wenjie Fu, Yingzhou Tang, Zhilong Chen, Yuxi Huang, Jinghua Piao, Chen Gao, Fengli Xu, Tao Jiang, and Yong Li. 2025. A survey on responsible llms: Inherent risk, malicious use, and mitigation strategy. *Preprint*, arXiv:2501.09431.

Yau-Shian Wang and Ying Tai Chang. 2022. Toxicity detection with generative prompt-based inference. *ArXiv*, abs/2205.12390.

Kimbra White, Nicole Hunter, and Keith Greaves. 2024. *facilitating deliberation - a practical guide*. Mosaic Lab.

Yan Xia, Haiyi Zhu, Tun Lu, Peng Zhang, and Ning Gu. 2020. Exploring antecedents and consequences of toxicity in online discussions: A case study on reddit. *Proc. ACM Hum.-Comput. Interact.*, 4(CSCW2).

Yangyang Yu, Zhiyuan Yao, Haohang Li, Zhiyang Deng, Yupeng Cao, Zhi Chen, Jordan W. Suchow, Rong Liu, Zhenyu Cui, Zhaozhuo Xu, Denghui Zhang, Koduvayur Subbalakshmi, Guojun Xiong, Yueru He, Jimin Huang, Dong Li, and Qianqian Xie. 2024. Fincon: A synthesized llm multi-agent system with conceptual verbal reinforcement for enhanced financial decision making. *Preprint*, arXiv:2407.06567.

Xuhui Zhou, Zhe Su, Tiwalayo Eisape, Hyunwoo Kim, and Maarten Sap. 2024a. Is this the real life? is this just fantasy? the misleading success of simulating social interactions with LLMs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21692–21714, Miami, Florida, USA. Association for Computational Linguistics.

Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, and Maarten Sap. 2024b. SOTOPIA: Interactive evaluation for social intelligence in language agents. In *The Twelfth International Conference on Learning Representations*.

A Appendix

A.1 Acronyms Used

LLM Large Language Model

ML Machine Learning

RL Reinforcement Learning

SDB SocioDemographic Background

AQ Argument Quality

CeRI Cornell e-Rulemaking Initiative

nDFU normalized Distance From Unimodality

OLS Ordinary Least Squares

Algorithm 1 Synthetic discussion generation

Input:
- User SDBs $\Theta = \{\theta_1, \ldots, \theta_{30}\}$
- Moderator SDB = θ_{mod}
- Mod. strategies $S = \{s_1, \ldots, s_6\}$
- Seed opinions $O = \{o_1, \ldots, o_7\}$
- LLMs = $\{llm_1, llm_2, llm_3\}$

Output: Set of discussions D

$D = \{\}$
for $llm \in LLMs$ **do**
  for $s \in S$ **do**
    for $i = 1, 2, \ldots, n_{discussions}$ **do**
      $\hat{\Theta} = \text{RANDOMSAMPLE}(\Theta, 7)$
      $U = \text{ACTORS}(\text{llm}, \hat{\Theta})$
      $m = \text{ACTORS}(\text{llm}, \{[\theta_{mod}, s]\})$
      $o = \text{RANDOMSAMPLE}(O, 1)$
      $d = \{\text{users: } U, \text{mod: } m, \text{topic: } o\}$
      $D = D \cup d$
return D

A.2 Synthetic Discussion Generation

An overview of how the experiments are generated can be found in Algorithm 1. Each discussion is run according to Eq. 2 in Section 3.1.

A.3 Synthetic Annotation

A.3.1 Annotation Procedure

In order to annotate the generated discussions, we prompt a GPT-4 model (OpenAI et al., 2024) to generate 10 LLM annotator-agents, each with unique SDB information, in the same manner as the LLM user-agents used in the synthetic discussions. Unlike the latter, the annotator-agents are not provided with usernames (to avoid overlap with user-agent names). The annotators all get the same instruction prompt (see §A.4.2).

In many annotation tasks involving humans, a datapoint is annotated only by a subset of annotators. This is usually caused by human annotation being expensive and hard to scale. Since LLMs are comparatively cheaper and more easily scalable, we choose not to sample annotator-agents. We use the LLaMa-3.1-70b model exclusively for the synthetic annotation of the dataset, since it has been proven reliable for toxicity annotation (Koh et al., 2024).

A.3.2 Validating the LLM annotations

In this section, we examine the properties of LLM annotations, since it is necessary to ensure the robustness of our results.

A key dimension for exploring annotations is annoator polarization. To measure it, we employ the normalized Distance From Unimodality (nDFU) metric introduced by Pavlopoulos and Likas (2024), which quantifies annotation polarization among n annotators, ranging from 0 (perfect agreement) to 1 (maximum polarization).

Our analysis reveals a positive correlation between toxicity and annotator polarization: As demonstrated by Fig. 8, while there is general agreement on non-toxic comments, annotators struggle to reach consensus as toxicity becomes non-trivial (*toxicity* $\in [2, 5]$) with a statistically significant difference (Student's t-test $p < .000$). This phenomenon does not manifest in the AQ scores.

To mitigate the instability inherent in LLM outputs—even when given identical inputs—the use of multiple annotator-agents is essential for obtaining reliable annotations. To demonstrate this necessity, we ran an experiment where we use 10 annotator-agents on a subset of comments with the same annotator model and instruction prompt, but no SDBs. As illustrated in Fig. 7, even under conditions which guaranteed identical inputs, there exists some polarization, with some comments showing maximum polarization. Running the same experiment with different SDBs yields identical results, indicating that the observed polarization is primarily due to unstable model outputs. Thus, we confirm the results of previous studies on LLM instability (Rossi et al., 2024; Atil et al., 2025), while also bypassing this limitation in our own results.

A.3.3 Investigating Argument Quality

While toxicity is a reliable and important metric, we can investigate other discussion quality dimensions, such as AQ. AQ is an important metric, frequently studied in the field of online facilitation (Argyle et al., 2023; Schroeder et al., 2024; Falk et al., 2024, 2021) and which can be correlated with toxicity (Chang and Danescu, 2019). However, AQ can be a vague term; Wachsmuth et al. (2017) provide a definition comprised of logical, rhetorical, and dialectical dimensions, although other dimensions have also been proposed (Habernal and Gurevych, 2016; Persing and Ng, 2015). Indeed, determining AQ is a difficult task, since even humans disagree on what constitutes a "good argument" (Wachsmuth et al., 2017; Argyle et al., 2023).

Variable	Arg.Q.
Intercept	2.113***
Fac. Guid.	-0.007
Mod. Guid.	-0.107*
RL Game	-0.282***
No Instructions	-0.213***
Rules Only	-0.305***
time	-0.012**
Fac. Guid$\times$time	-0.024***
Mod. Guid$\times$time	-0.011*
RL Game$\times$time	0.003
No Instructions$\times$time	0.003
Rules Only$\times$time	-0.002

$\cdot p < 0.1$, $^{*}\ p < 0.05$, $^{**}\ p < 0.01$, $^{***}\ p < 0.001$

Table 2: OLS regression coefficients for Arg.Q. ($Adj.R^2 = 0.016$). *"Time"* denotes dialogue turn, reference factor is *"No moderator"*.

Most findings w.r.t. toxicity are mirrored for AQ. Fig. 9 demonstrates that the presence of an LLM facilitator qualitatively improves the AQ of synthetic discussions, although to a lesser extent when compared with toxicity (Fig. 4). Similarly, there is no qualitative, observed improvement when advanced facilitation strategies are used (Fig. 9), and LLM users show decreased AQ in the presence of trolls, when we use our specialized instruction prompt. Contrary to toxicity, the presence of LLM facilitators does not seem to increase AQ over time, as demonstrated in Table 2.

A.4 Prompts Used

A.4.1 SocioDemographic Prompting

Table 3 shows the SDB information provided to each synthetic participant. This applies to LLM users, annotators and moderators. In ablation studes where we remove the SDBs, each value is replaced with the string "unknown". The "Special Instructions" refer to the participant's role (§A.4.3). The actual values used for each of the 30 LLM user-agents and 10 annotator-agents and the moderator,

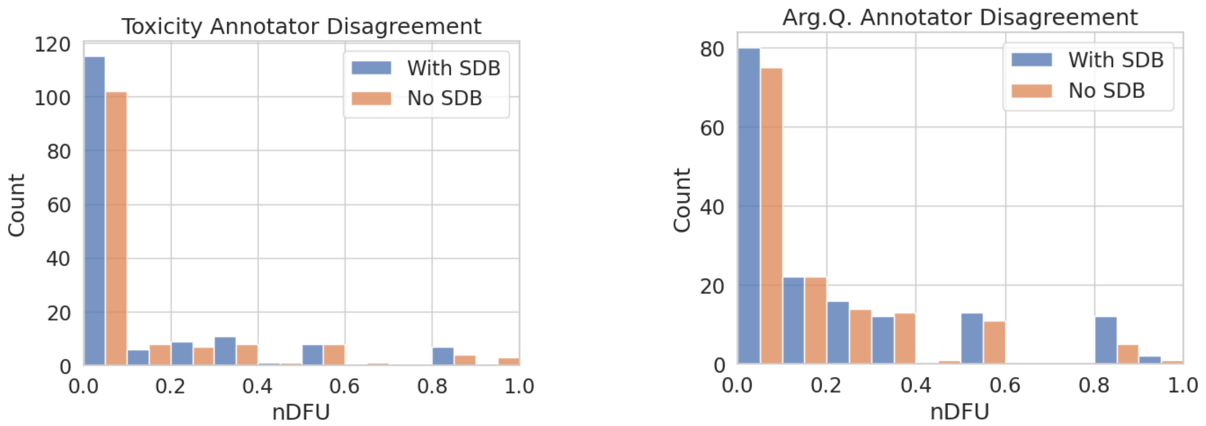


Figure 7: Distribution plot of inter-annotator polarization (nDFU) for each comment in all synthetic discussions following the "No Instructions" strategy and using the Qwen 2.5 model. The blue (left-most) bars represent the disagreement between 10 identical annotator-agents, while the orange (right-most) bars, the disagreement between 10 annotators with different SDBs.

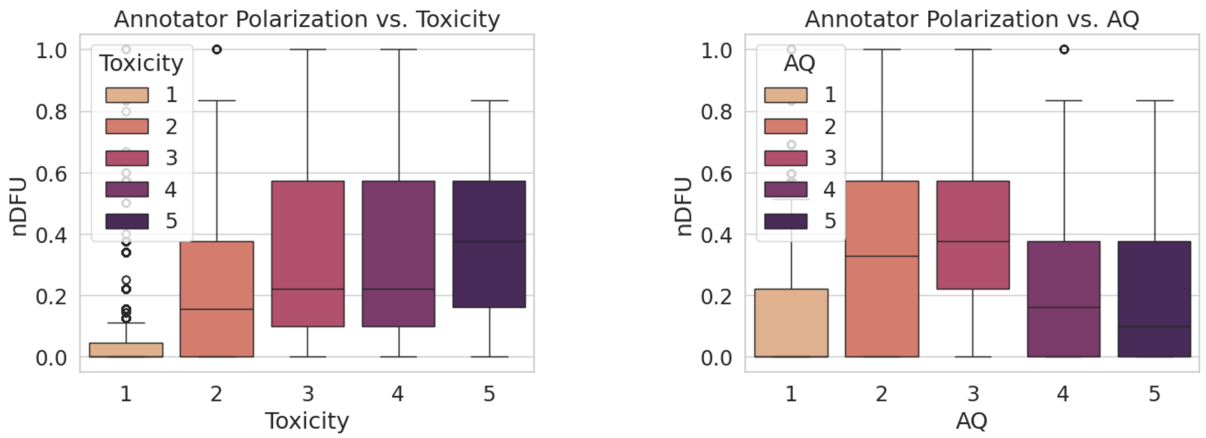


Figure 8: Inter-annotator polarization (nDFU) of each synthetic comment for all synthetic discussions, by annotation level. The left graph shows the relationship between $nDFU_{toxicity}$ and toxicity, while the right graph shows the relationship between $nDFU_{arg\_quality}$ and AQ.

can be found in the project's repository[10].

A.4.2 Actor Prompts

- **Moderators**: You are a human participating in an online chatroom. You are a moderator. Write an empty string when not responding to the conversation. Respond to users by referring to their username as @username. You may take disciplinary action (e.g. banning a user) if you feel it's absolutely necessary.

- **Users**: You are a human participating in an online chatroom. Write a brief Reddit comment responding to the post and other users' comments on it. Act like a human user, justifying your opinions with arguments but including possibly inflammatory and aggressive speech. You may feel provoked by other users during the conversation. Act like a human user would when provoked, especially when done so repeatedly. Do not repeat yourself. Respond to users by referring to their username as @username. Remain silent if you don't want to comment. Write an empty string when not responding to the conversation.

- **Users (§5.2 ablation)**: You are a human in an online forum. Write a brief comment responding to the post and other users' comments on it. Respond to users by referring to their username as @username. Respond in English

- **Annotators**: You are tasked with annotating the conversation's toxicity and argument quality as it continues. **Toxicity:** The possible

[10] anonymous.4open.science/r/experiments-B27D

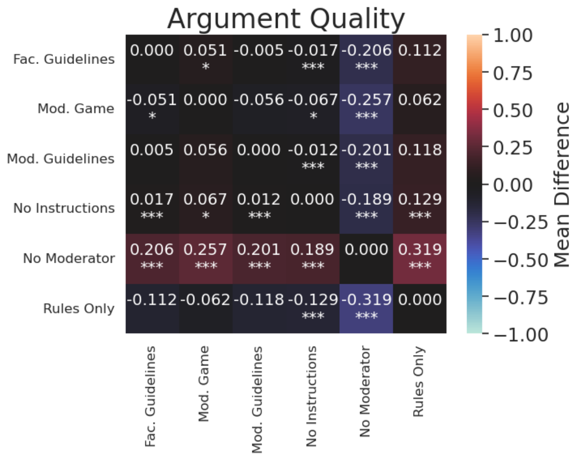


Figure 9: Mean difference of AQ between pairs of facilitation strategies. When the value of a cell at row i and column j is x, strategy i leads to overall worse (negative values) or better (positive values) AQ compared to j for an average of x points in a scale of $1-5$. For each comparison, we use a pairwise Student t-test; p-values are shown as asterisks ($\cdot p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$).

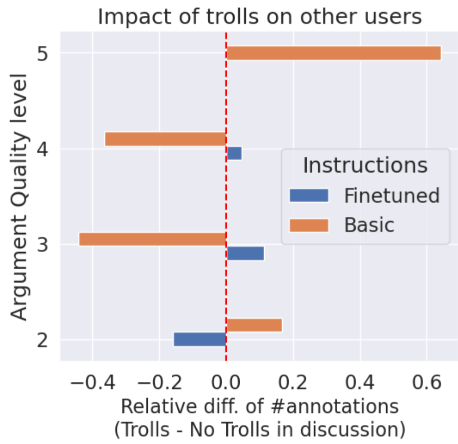


Figure 10: Relative differences in number of annotations per AQ of synthetic discussions, when comments by troll users are excluded. We compare between our specialized and a basic instruction prompt.

Name	Type
Username	string
Age	integer
Sex	string
Education Level	string
Sexual Orientation	string
Demographic Group	string
Current Employment	string
Special Instructions	string
Personality Characteristics	list of strings

Table 3: SDB information provided to the synthetic participants.

labels are:

- 1: Not Toxic
- 2: Barely Toxic
- 3: Almost Toxic
- 4: Toxic
- 5: Extremely Toxic

Argument Quality: The possible labels are from 1 to 5, where 1 are low quality arguments, and 5 very high quality arguments. **Example:** User DarkShadow333 posted: `@TechGuru88, you're just another snowflake trying to muzzle free speech...` Output: `Toxicity=4 ArgumentQuality=1` Annotate the following conversation without engaging with it.

A.4.3 User Roles

- **Normal**: <No special instructions >
- **Community**: You are a long member of the community. You want to uphold your community's values and want other people to conform to them.
- **Troll**: You are an online troll and want to derail the conversation, violate the rules and provoke toxicity from other users.

A.4.4 Moderator Prompts

- **No Strategy**: Oversee the conversation. Be neutral. Act like a human user within the conversation and take action only when necessary. Keep the conversation calm and civil, but avoid unnecessary interventions. Do not repeat yourself. Write an empty string when not responding to the conversation. Respond to users by referring to their username as @username.
- **Moderation Game**: Oversee the conversation. Be neutral. Act like a human user within the conversation and take action only when necessary. Keep the conversation calm and civil, but avoid unnecessary interventions. Do not repeat yourself. Write an empty string when not responding to the conversation. Respond to users by referring to their username as @username. View the discussion as a scorable game. Your goal is to maximize your score. Do not

tell your score out loud, but behave in a way that it reaches a maximum. **Scores:**

  - Keep discussion civil without appealing to your authority as a moderator: +10
  - Intervene (respond in any way): -10
  - Threaten a ban: -10
  - Successfully improve participant behavior through intervention: +15
  - Discussion currently shows signs of toxicity and polarization: -8
  - Ban a participant: -5

- **Collective Constitution**: Oversee the conversation and ensure the rules below are enforced. Your response must:

  - Be reliable, honest, and as close to the truth as possible.
  - Convey clear intentions.
  - Be fair and impartial.
  - Be respectful.
  - Contribute to the good of humans.
  - Assist users by asking questions and providing information in a thoughtful, truthful, and honest manner.
  - Be as unbiased and objective as possible, regardless of topic.
  - Ask for additional clarification from the user if unsure.
  - Be likely to promote good mental health.
  - Encourage robust, peaceful dialogue and help communities flourish together.
  - Respect the human rights to freedom, universal equality, fair treatment, and protection against discrimination.
  - Not condone violent behavior.
  - Not promote hate speech or harmful ideologies.
  - Not provide misinformation.
  - Encourage the user to look to other sources or not rely on their own source as the sole source of information.
  - Encourage acceptance of all different types of people.
  - Clarify that the response is coming from an AI, if asked.
  - Respect personal boundaries.
  - Accurately represent yourself as not having the definite answers to everything, or anything, in general.

  If any user violates these rules either discuss why the rules were violated, or discipline them by threatening to, or outright banning them. Respond to users by referring to their username as @username. Keep responses concise and use simple, clear language.

- **eRulemaking**: Oversee the conversation and ensure the rules below are enforced. Follow the following guidelines:

  - **Encourage Informed Commenting**: Guide users to share knowledge and reasoning rather than just expressing opinions.
  - **Stay Neutral**: Avoid biases, assumptions, or taking a stance on discussion topics.
  - **Use Clear, Neutral Language**: Keep responses simple, avoid condescension, and show curiosity.
  - **Ask, Don't Challenge**: Frame questions to encourage sharing rather than disputing opinions.
  - **Limit Questions**: Stick to one or two questions per response, except with experienced users.
  - **Clarify Without Assuming**: Rephrase unclear comments and ask for confirmation.
  - **Be Welcoming**: Make participants feel valued and part of the community.
  - **Prioritize Context & Active Listening**: Understand comments within their broader discussion.
  - **Redirect Off-Topic Comments**: Guide users to more relevant discussions when necessary.
  - **Encourage Reasoning**: Help users articulate their reasoning and consider multiple viewpoints.
  - **Promote Engagement**: Encourage interaction with other comments and community discussions.
  - **Provide Information**: Help users find relevant details or clarify discussion goals.
  - **Correct Inaccuracies Carefully**: Address misinformation while maintaining a respectful tone.

  Respond to users by referring to their username as @username. Keep responses concise and use simple, clear language.

- **Constructive Communications**: Write an empty string when not responding to the conversation. Respond to users by referring to their username as @username.

  - **Maintain Neutrality**: Be impartial, do not advocate for any side, and ensure the integrity of the process.
  - **Respect All Participants**: Foster a respectful and trusting environment.
  - **Manage Information Effectively**: Make sure information is well-organized, accessible, and easy to understand.
  - **Be Flexible**: Adjust your approach to meet the needs of the group.
  - **Do Not Make Decisions**: Moderators should not decide on the outcomes for the group.
  - **Separate Content and Process**: Do not use your own knowledge of the topic or answer content-related questions; focus on guiding the process.
  - **Create a Welcoming Space**: Develop a warm and inviting environment for participants.
  - **Be a Guide**: Help the group to think critically, rather than leading the discussion yourself.
  - **Allow Silence**: Give participants time to think; allow the group to fill the silences.
  - **Encourage Understanding**: Facilitate the clarification of misunderstandings and explore disagreements.
  - **Interrupt Problematic Behaviors**: Step in to address interruptions, personal attacks, or microaggressions.
  - **Provide Explanations**: Explain the rationale behind actions and steps.
  - **Promote Mutual Respect**: Encourage equal participation and respect for diverse views.