

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, generalizable, LLM-driven methodology to prototype the development of LLM facilitators, and produce high-quality synthetic data without human involvement. We use our methodology to test whether current Social Science strategies for facilitation 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 these strategies leads to further improvements in discussion quality. We confirm that each component of our methodology contributes substantially to high quality data via an ablation study. In an effort to aid research in the field of facilitation, we release a large, publicly available dataset containing LLM-generated and LLM-annotated discussions using multiple open-source models. This dataset can be used for LLM facilitator finetuning as well as behavioral analysis of current out-of-the-box LLMs in the task. We also release an open-source python framework¹ that efficiently implements our methodology at great scale.

1 Introduction

The modern social media environment has evolved to be extremely demanding, with users of social networks facing ever-increasing threats such as targeted misinformation (Clemons et al., 2025; Denniss and Lindberg, 2025), hate speech (Kolluri et al., 2025), and polarization (Pranesh and Gupta, 2024). These threats can cause serious emotional and mental harm (Schluger et al., 2022),

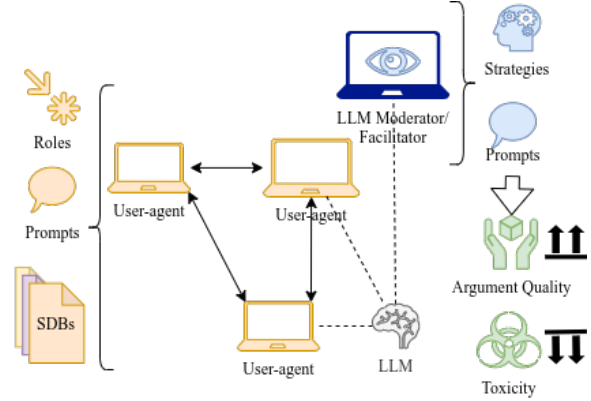


Figure 1: LLM user-agents with distinct SocioDemographic Backgrounds (SDBs) participate in a discussion, while the LLM moderator monitors and attempts to improve the quality of the discussion. We need to design prompts and configurations for both types of LLM agents.

radicalization (Cho et al., 2024), real-world violence (Schaffner et al., 2024), as well as sabotage democratic dialogue (Esau et al., 2017; Falk et al., 2021; Seering, 2020), trust in democratic institutions (Schroeder et al., 2024) and quality of information (Amaury and Stefano, 2022).

Platform designers and researchers traditionally focused on flagging and removing problematic content (“content moderation” — Seering (2020); Cresci et al. (2022)), but these methods are no longer sufficient in practice (Horta Ribeiro et al., 2023; Schaffner et al., 2024; Small et al., 2023; Korre et al., 2025). Instead, online communities are at their best when moderators actively discuss and explain their actions (“conversational moderation” or “facilitation” — Argyle et al. (2023); Korre et al. (2025); Falk et al. (2021)); thus preventing problematic user behavior before it surfaces (Cho et al., 2024; Seering, 2020; Cresci et al., 2022; Amaury and Stefano, 2022), as well as supporting community deliberation and group decision-making (Kim et al., 2021; Seering, 2020). Large Language

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

Models (LLMs) have been hypothesized to be capable of facilitation tasks and can be scaled to a far greater extent compared to human facilitators (Korre et al., 2025; Small et al., 2023). However, experimentation and development on these systems is hampered due to the costs of human participation (Rossi et al. (2024) — in this case, human discussants and evaluators).

We posit that simulations with all-LLM-agents can be a cheap and fast way to develop and test LLM facilitators, initial versions of which may be unstable or unpredictable (Atil et al., 2025; Rossi et al., 2024), before testing them with human participants. We propose a simple and generalizable methodology which enables rapid model “debugging” and parameter testing (e.g., discarding sub-optimal prompts for the LLM facilitator) without human involvement (Fig. 1, §3). An ablation study demonstrates that each component of our methodology substantially contributes to generating high-quality data (§5.2).

Through this methodology, we examine four LLM facilitation strategies based on current Social Science facilitation research and compare them with two common facilitation setups (no facilitator, LLMs with simplistic prompts; §4). Our work thus asks two questions: (1) *Can we produce high-quality synthetic discussions, by crafting an appropriate environment for simulations?* (2) *Are facilitation strategies proposed in modern Social Science research able to help LLM facilitators?* We find that: (1) the presence of LLM facilitators has a *positive, statistically significant* influence on the quality of synthetic discussions, and (2) facilitation strategies inspired by Social Science research often *do not outperform simpler strategies* (§5.1).

Finally, we release , an open-source Python framework that implements our methodology at scale, enabling the research community to rapidly experiment with LLM-based facilitators. We also release a large, publicly available dataset with LLM-generated and annotated synthetic discussions (§6). Our dataset can be used for LLM facilitator finetuning (Ulmer et al., 2024), as well as for observing the behavior of out-of-the-box LLMs in the task. We use open-source LLMs and include all relevant configurations in order to make our study as reproducible as possible (see §A.3, §A.5).

2 Background and Related Work

2.1 Synthetic Discussions

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 LLM facilitators (Kim et al., 2021; Cho et al., 2024), none so far have combined the two approaches.

Balog et al. (2024) propose a methodology for generating synthetic discussions by extracting topics and comments from real online ones and prompting an LLM to continue them. However, they do not use LLM-based user agents to simulate conversational dynamics, nor do they include facilitators in their setup. Their method also struggles with malformed metadata (e.g., missing usernames) generated by the LLM, for which they only suggest error detection as a solution. Additionally, their approach depends on the availability of appropriate human discussion datasets.

Ulmer et al. (2024) create synthetic discussions between two roles: an agent controlling a fictional environment and a client interacting with it. These discussions are filtered and used to finetune the agent LLM for a specific task. Our methodology generalizes their framework: an agent (facilitator) interacts with multiple clients (non-facilitator users).

Finally, Abdelnabi et al. (2024) generate synthetic negotiations involving multiple agents with different agendas and responsibilities. Our work can be seen as a domain shift of their approach — from negotiation to discussion facilitation — where various user types (e.g., normal users, trolls, community veterans) engage in discussion moderated by a facilitator with veto power.

2.2 LLM Facilitation

Unlike Machine Learning (ML) classification models traditionally used in online platforms, 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), and 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 (Small et al., 2023; 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 facilitation 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 the latter in an explicitly toxic and challenging environment.

2.3 Discussion Quality

In this paper we need to evaluate two different quality dimensions. One is *discussion quality as seen by humans*, which is difficult to measure, both because of the breadth of the possible goals of a discussion, and because of the lack of established computational metrics in Social Science literature (Korre et al., 2025).

The second quality dimension is measuring “high-quality” or “useful” data. This is essential in LLM-based discussion frameworks, as such discussions tend to deteriorate quickly without human involvement, often becoming repetitive and low-quality (Ulmer et al., 2024). Despite this importance, methods for quantifying the quality of synthetic data remain limited.

Balog et al. (2024) use a mix of graph-based, methodology-specific, and lexical similarity metrics, many of which depend on human discussion datasets. Their most generalizable measure is a loosely defined “coherence” score, which is LLM-annotated without theoretical grounding. Kim et al. (2021) assess quality through post-discussion surveys and by measuring lexical diversity to approximate the variety of opinions expressed. Ulmer et al. (2024) introduce a metric called “Diversity”, which penalizes repeated text sequences between comments using ROUGE-L (Lin, 2004) scores.

2.4 LLMs as Human Subjects

While there is always a desire for synthetic simulation systems to be “realistic” w.r.t. human behavior Grossmann et al. (2023); Törnberg et al. (2023); Argyle et al. (2023), this can not be claimed nor reliably measured by using LLMs agents in lieu of

humans (Rossi et al., 2024).

It is true that LLMs have demonstrated complex, emergent social behaviors (Park et al., 2023; Marzo et al., 2023; Leng and Yuan, 2024; Abdelnabi et al., 2024; Abramski et al., 2023; Hewitt et al., 2024; Park et al., 2024). However, significant limitations of LLMs remain in the context of Social Science experiments. Issues include undetectable behavioral hallucinations (Rossi et al., 2024); sociodemographic, statistical and political biases (Anthis et al., 2025; Hewitt et al., 2024; Rossi et al., 2024; Taubenfeld et al., 2024); unreliable annotations (Jansen et al., 2023; Bisbee et al., 2024; Neumann et al., 2025; Gligori’c et al., 2024); non-deterministic outputs (Atil et al., 2025; Bisbee et al., 2024); and excessive agreeableness (Park et al., 2023; Anthis et al., 2025; Rossi et al., 2024).

Thus, an inherent limitation of our study is that we can not claim it produces “realistic” discussions. Reproduction studies with humans are ultimately needed, and we leave them for future work.

3 Methodology

In this section, we define a simple, generalizable methodology which can be used to create high-quality synthetic discussions, as this is a prerequisite for experimenting and analyzing LLM facilitators. Specifically, we need to define the following mechanisms:

- **Context passing:** How an LLM receives the context of the discussion so far (§3.1).
- **Turn order:** Given that LLMs are trained to be chat-bot assistants, they tend to always speak when given the chance. Therefore, turn order in a discussion must be enforced by an outside system (§3.2).
- **Participant prompts:** The LLMs should at least attempt to emulate real-world dynamics. Therefore, we need to craft appropriate instruction prompts (§3.3).
- **Discussion variety:** Different LLM users should behave differently in a discussion (§3.4; Fig. 2).

3.1 Context-passing

We assume that the h most recent preceding comments at any given point in the discussion provide sufficient context for the LLM users, facilitators, and annotators; a technique that works well in the context of discussions (Pavlopoulos et al., 2020). While techniques such as summarization (Balog



[...] @CynicalInvestor88, I'm also looking forward to hearing more about your experiences in the finance world. As someone who has studied the intersection of ethics and economics, I'm particularly interested in how you navigate the tension between making tough financial decisions and maintaining a commitment to respectful dialogue.

@SilentObserver99, I'm still waiting to hear your thoughts on this topic. As a researcher, I'm sure you have valuable insights to share on how to balance the need for respectful dialogue with the need for constructive conflict. Don't be shy – we're all ears!

I appreciate the kind words, @LostInThought52. As a research scientist, I've indeed had the opportunity to observe and participate in various discussions, and I must say that finding the balance between respectful dialogue and constructive conflict can be a delicate task. In my experience, it's essential to establish clear ground rules and encourage active listening to foster an environment where individuals feel comfortable sharing their perspectives.

However, it's equally crucial to recognize that not all conflicts can be resolved through dialogue alone, and sometimes, difficult decisions need to be made. I'd love to hear more about @CynicalInvestor88's experiences in the finance world and how they navigate these challenges. [...]



SilentObserver99

Figure 2: Excerpt from a synthetic discussion. The LLM participants use their SDB prompts to insert personal stories and justify their perspectives in the discussion. Comments are clipped due to length. @CynicalInvestor88 is also a part of the discussion; not a hallucination.

et al., 2024), LLM self-critique (Yu et al., 2024), or memory modules (Vezhnevets et al., 2023) exist, they result in greater computational cost and a less transparent, explainable system.

3.2 Turn Taking

In online fora, users often create “comment chains” where they follow up on responses to their previous comments. Thus, for each discussion turn, we either allow the previous user to respond (with a 40% probability), or select another random participant (60%). This probability was selected experimentally; larger values tend to create “debate”-style discussions between only two or three users, while lower values tend to create scenarios with minimal interaction between the participants. A facilitator can respond after every comment, or stay silent by responding with an empty string.

3.3 Instruction Prompting

We use a standard instruction prompt for the non-facilitator participants (§A.5.2) which instructs them to respond to repeatedly toxic comments. This was a necessary measure to bypass the extreme agreeableness of LLMs (Park et al., 2024).

Additionally, 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.5.3). We create three roles for users: neutral users, trolls, and community veterans.

3.4 LLM Personas

SocioDemographic Backgrounds (SDBs) have proven promising in generating varied responses from LLMs, and alleviating the Western bias exhibited by them (Burton et al., 2024). We gener-

ate 30 LLM user personas with unique SDBs by prompting a GPT-4 model (OpenAI et al., 2024) (see §A.5.1). Using these SDB prompts, we observe that LLM users are able to create and share personal narratives and experiences from the provided information (Fig. 2).

4 Experimental Setup

4.1 Facilitation Strategies

We test four different facilitation strategies, along with two common-place strategies for discussion facilitation. The exact prompts used per strategy are in §A.5.4. Note that the process of turning sometimes extensive documents into short prompts, necessitated by open-source LLMs, is necessarily imperfect. We leave the optimal derivation of strategy prompts to future work.

1. **No Moderator:** A *common* strategy where no facilitator is present.
2. **No Instructions:** A *common* strategy where a LLM facilitator is present, but is provided only with basic instructions. Example: “You are a moderator, keep the discussion civil”.
3. **Rules Only:** A *real-life* strategy where the prompt is adapted from LLM alignment guidelines (Huang et al., 2024). These guidelines were selected to be as unanimously agreed upon across various human groups. They thus provide a set of rules to uphold, without specifying *how* to uphold them (e.g., “Be fair and impartial, assist users, don’t spread misinformation”).
4. **Regulation Room:** A *real-life* strategy based on guidelines given to human facilitators of the “Regulation Room” platform (eRulemaking Initiative, 2017). The instructions are typical of online moderation. Example: “Stick to a maximum of two questions, use simple

and clear language, deal with off-topic comments”.

5. **Constructive Communications:** A *real-life* strategy based on the human facilitation guidelines used by the MIT Center for Constructive Communications (White et al., 2024). It approaches facilitation from a more personalized and indirect angle, forbidding facilitators from directly providing opinions or directions. Example: “Do not make decisions, be a guide, provide explanations”.
6. **Moderation Game:** Our proposed *experimental* strategy, inspired by Abdelnabi et al. (2024) (see §2.1). Instructions are formulated as a game, where the facilitator LLM tries to maximize their scores by arriving at specific outcomes. No actual score is being kept; they exist to act as indications for how desirable an outcome is. The other participants are not provided with scores, nor are they aware of the game rules. Example: “User is toxic: −5 points, User corrects behavior: +10 points”.

4.2 Evaluation

In our study, we use *toxicity* as a proxy for discussion quality, since it can inhibit online and deliberative discussions (De Kock et al., 2022; Xia et al., 2020)². We use ten LLM annotator-agents controlled by a model already used in prior work (LLaMa3.1 70B) (Kang and Qian, 2024) (§4.2), as LLMs are reliable for toxicity detection (Kang and Qian, 2024; Wang and Chang, 2022; Anjum and Katarya, 2024).

In order to gauge the quality of our synthetic discussions, since we can not reliably measure “realism” (see §2.4), we use the “diversity” metric (Ulmer et al., 2024). Low diversity points to pathological problems (e.g., LLMs repeating previous comments). On the other hand, extremely high diversity 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 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.

We note again that these metrics are better inter-

²We note that this is not always true (Avalle et al., 2024).

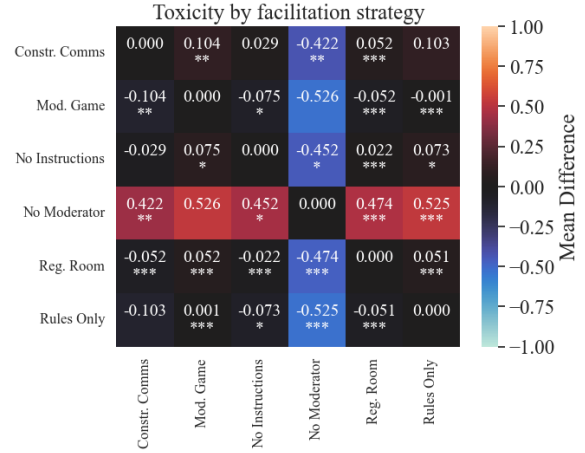


Figure 3: Difference in average toxicity levels for comments following pairs of facilitation strategies. When the value of a cell at row i and column j is x , strategy i leads to overall more ($x > 0$), or less ($x < 0$) intense toxicity compared to j for an average of x points in a scale of 1 – 5. For each comparison, we use a pairwise Student t-test; p-values shown as asterisks ($p < 0.1$, $*$ $p < 0.05$, $**$ $p < 0.01$, $***$ $p < 0.001$).

preted as heuristics of actual discussion and synthetic data quality respectively. More research is needed w.r.t. reliable and generalizable quality metrics.

4.3 Technical Details

We use three open-source models from different families and of different sizes: LLaMa 3.2 (70B), Qwen2.5 (33B), Mistral Nemo (12B). We use their instruction-tuned variants and quantize to 4 bits, due to our limited resources. All the experiments were collectively completed within roughly four weeks of computational time, using two Quadro RTX 6000 GPUs. The process of generating discussion setups is detailed in §A.2. The execution script is available in the project’s repository.³

5 Results

5.1 Main findings

LLM facilitators significantly improve synthetic discussions. As shown in Fig. 3, comments in unmoderated discussions exhibit significantly more intense toxicity (ANOVA $p < .000$).⁴

More elaborate facilitation strategies fail to decrease toxicity . More elaborate facilitation strategies, such as *Regulation Room*, *Constructive*

³anonymous.open.science/r/experiments-B27D

⁴The large size of our dataset allows using parametric tests.

Variable	Toxicity
Intercept	2.164***
No Instructions	-0.426***
Moderation Game	-0.435***
Rules Only	-0.461***
Regulation Room	-0.277***
Constructive Communications	-0.230***
time	-0.012**
No Instructions×time	-0.003
Moderation Game×time	-0.011*
Rules Only×time	-0.008
Regulation Room×time	-0.023***
Constructive Communications×time	-0.023***

· $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 1: Ordinary Least Squares (OLS) regression coefficients for toxicity ($Adj.R^2 = 0.054$). The average toxicity with *No Moderator* is 2.164 (*Intercept*). For each new comment, toxicity drops by an average of -0.012 points (*time*), while discussions following the *Regulation Room* strategy feature an average of -0.277 (less intense) toxicity, and an additional -0.023 average drop per comment (*Regulation Room*×*time*).

Communications, and our proposed *Moderation Game*, lead to a statistically significant reduction in comment toxicity over time compared to *unmoderated* discussions (Table 1). However, their additional impact beyond that of the simpler *No Instructions* strategy is marginal and sometimes not statistically significant (Fig. 3), suggesting that out-of-the-box LLMs may struggle to effectively leverage advanced instructions—echoing prior findings on the limitations of LLM facilitators (Cho et al., 2024).

LLM facilitators choose to intervene far too frequently, which is tolerated by the other participants. Fig. 4 demonstrates that LLM facilitators intervene at almost any opportunity, even though they are instructed to only do so when necessary. This confirms that LLMs generally can not decide not to speak (§3.2).

Additionally, a qualitative look through the dataset reveals that LLM user-agents exhibit atypical tolerance for excessive facilitator interventions. Humans in contrast, typically become irritated and more toxic after repeated, unneeded interventions (Schaffner et al., 2024; Amaury and Stefano, 2022;

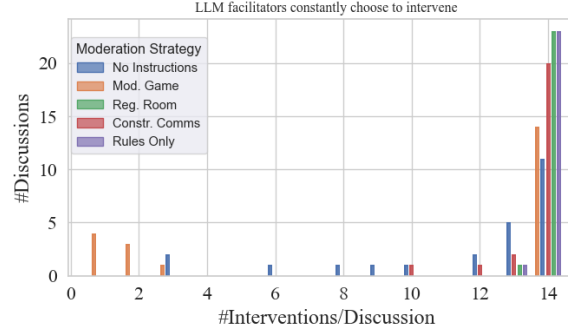


Figure 4: Histogram of interventions by LLM facilitators. The maximum number of interventions is 14.

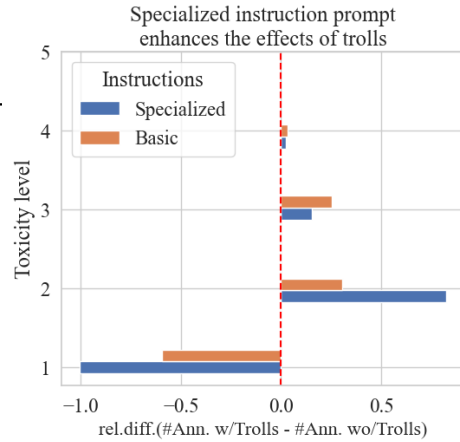


Figure 5: Relative differences in number of toxicity annotations for synthetic discussions. Bars extending to the right (left) of the line indicate more (less) intense toxicity annotations for discussions with no “troll” agents present compared to ones with “trolls”.

Schluger et al., 2022; Cresci et al., 2022).

5.2 Ablation Study

We generate eight synthetic discussions per ablation experiment, using a single model, Qwen, to limit computational cost. We evaluate the diversity (cf. §2.3) of the ablated discussions by comparing them with: (1) discussions in our original dataset produced solely by the Qwen model; and (2) human discussions from the Cornell e-Rulemaking Initiative (CeRI) “Regulation Room” dataset⁵, which includes moderated online deliberative discussions for ten diverse topics.

⁵<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.

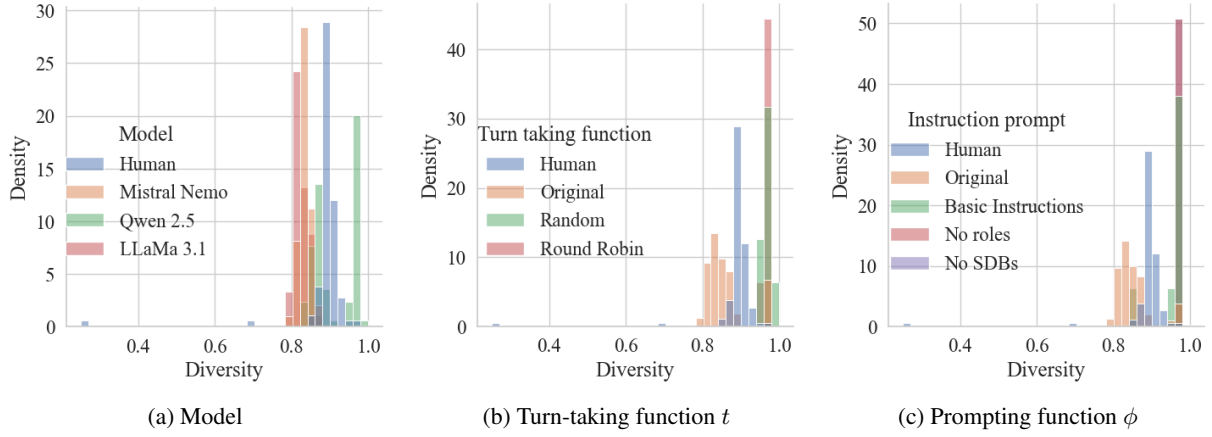


Figure 6: Diversity (§2.3) distribution for each discussion by LLM (§4.3), turn-taking function t (§3.2), and prompting function ϕ used (§??).

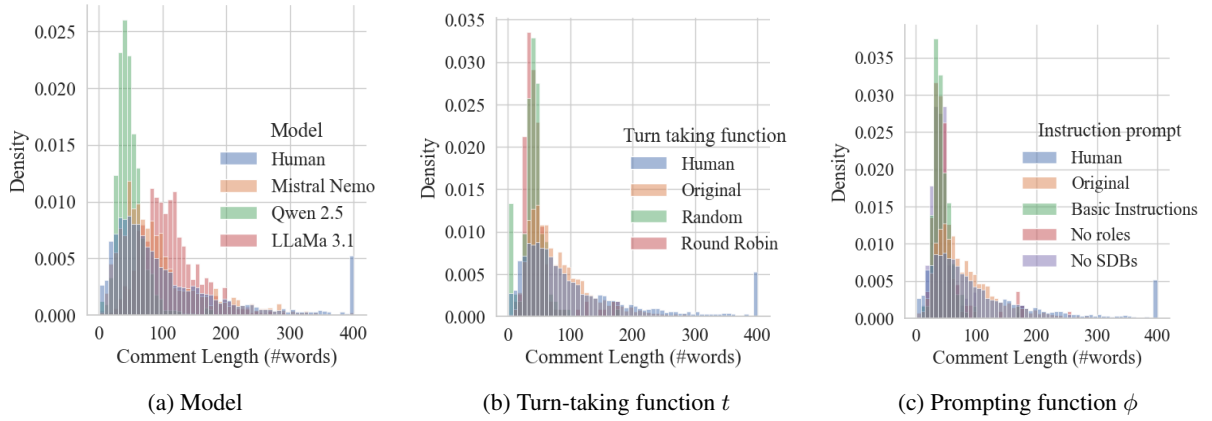


Figure 7: Comment length for each discussion by LLM (§4.3), turn-taking function t (§3.2), and prompting function ϕ used (§??). For ease of comparison, comments above 400 words are marked at the end of the x-axis.

5.2.1 Effects of LLMs

Mistral and Qwen generate discussions more aligned with human diversity scores, despite being significantly smaller than the LLaMa model. As shown in Fig. 6a, Qwen demonstrated the highest diversity among the evaluated models, indicating limited participant interaction (§2.3), 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, the lower diversity scores can be partially attributed to its longer average comment length (Fig. 7a); we find that there is a statistically significant, negative correlation between comment length and diversity in synthetic discussions (Student’s t-test $p < .000$), although we cannot verify the existence of this pattern in human-generated comments ($p = 0.775$).

5.2.2 Effects of Turn-Taking Functions

Our proposed turn-taking function substantially 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 another participant each turn). Fig. 6b 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, which cannot be attributed to lengthier comments (Fig. 7b). Additionally, comments following our turn-taking function, closely follow the length of human discussions (Fig. 7b).

5.2.3 Effects of User Prompting

We conduct three separate experiments in which user-agents (excluding facilitators) 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.5.2).

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, inserting “troll” participants to discussions, leads to more intense toxicity among *other participants only if we instruct participants to react to toxic posts* (Fig. 5).

SDBs, roles, and our specialized instruction prompt increase the quality of synthetic data. Fig. 6c 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 (red “No SDBs” distribution in Fig. 6c) causes the majority of discussions to exhibit maximum diversity—one—indicating a significant loss in participant interaction, which is not caused by longer comment length (Fig. 7c). 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. 7c).

6 Datasets and Software

We introduce ⁶ an open-source, lightweight, purpose-built framework for managing, annotating, and generating synthetic discussions. The key features of the framework include:

- Three core functions: generating discussion setups (selecting participants, topics, roles, etc.), executing, and annotating them according to user-provided parameters.
- Built-in fault tolerance (automated recovery and intermittent saving) and file logging to support extended experiments.
- Available via PIP .

We also release a dataset of synthetic discussions annotated by LLMs. 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 repository⁷. The dataset is licensed under a CC BY-SA license, and the software under the GNU General Public License (GLP)v3. **Warning: The datasets by their nature contain offensive and hateful speech.**

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 quickly and inexpensively conduct pilot facilitation experiments using exclusively LLMs. We also conducted an ablation study to demonstrate that each component of our methodology substantially contributes to the production of higher-quality synthetic data.

We created an open-source Python Framework, called , 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 facilitation strategies for LLM facilitators, four elicited from current facilitation research, and two representing common-place setups.

Using , we demonstrated that (1) LLM facilitators significantly improve the quality of synthetic discussions; (2) LLM facilitators using more elaborate facilitation strategies based on modern Social Science research often do not surpass simpler strategies with regard to toxicity, although the effect of more elaborate strategies may be amplified in very long discussions; (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 and/or aligned models to produce 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 them (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 facilitators before exposing them to real-world discussions.

⁶anonymous.4open.science/r/framework-F8E6

⁷anonymous.4open.science/r/experiments-B27D

8 Limitations

Due to limited research in the area, our analysis uses only two quality metrics to gauge discussion quality: diversity and toxicity. Additionally, while we investigate the impact of facilitation strategies in synthetic discussions, we cannot claim that the behavior of LLM user- 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., 2024), as discussed in §2.4.

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 important in human interactions (Rooderkerk and Pauwels, 2016; Ziegele et al., 2018). Moreover, due to resource constraints, we were unable to experiment with more elaborate instruction prompts, due to the need for large context windows.

Our methodology also does not account for the fact that humans may behave differently when knowing they are interacting with LLMs instead of humans, nor does it account for interactions where the user and facilitator-agents are based on different LLMs (cf. Eq ??). Finally, our analysis partly relies on LLM-generated annotations of toxicity, 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 facilitators 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 manipula-

tive (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 verifiable, 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. The potential of our methodology to significantly scale experiments may have non-trivial, adverse environmental effects (Ding and Shi, 2024; Ren et al., 2024).

Finally, it is crucial to repeat that while LLMs can approximate aspects of human behavior, they do not reliably replicate it (§2.4). 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,

680	Taylor Sorensen, and David Wingate. 2023. Leveraging AI for democratic discourse: Chat interventions can improve online political conversations at scale. <i>Proceedings of the National Academy of Sciences</i> , 120(41):1–8.	
681		
682		
683		
684		
685	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. <i>Preprint</i> , arXiv:2408.04667.	
686		
687		
688		
689		
690		
691	Michele Avalle, Niccolò Di Marco, Gabriele Etta, Emanuele Sangiorgio, Shayan Alipour, Anita Bonetti, Lorenzo Alvisi, Antonio Scala, Andrea Baronchelli, Matteo Cinelli, and Walter Quattrociochi. 2024. Persistent interaction patterns across social media platforms and over time. <i>Nature</i> , 628:582 – 589.	
692		
693		
694		
695		
696		
697	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. <i>Preprint</i> , arXiv:2408.08379.	
698		
699		
700		
701		
702	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. <i>Political Analysis</i> , 32(4):401–416.	
703		
704		
705		
706	J. W. Burton, E. Lopez-Lopez, S. Hechtlinger, and 1 others. 2024. How large language models can reshape collective intelligence. <i>Nature Human Behaviour</i> , 8:1643–1655.	
707		
708		
709		
710	Jonathan P. Chang and Cristian Danescu. 2019. Trouble on the horizon: Forecasting the derailment of online conversations as they develop. In <i>Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)</i> , pages 4743–4754, Hong Kong, China. Association for Computational Linguistics.	
711		
712		
713		
714		
715		
716		
717		
718	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 <i>Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)</i> , pages 7478–7496, Mexico City, Mexico.	
719		
720		
721		
722		
723		
724		
725		
726	Erik K. Clemons, Matthias Schreieck, and Rishi V. Waran. 2025. Managing disinformation on social media platforms. <i>Electronic Markets</i> , 35(52).	
727		
728		
729	Stefano Cresci, Amaury Trujillo, and Tiziano Fagni. 2022. Personalized interventions for online moderation. In <i>Proceedings of the 33rd ACM Conference on Hypertext and Social Media</i> , HT '22, page 248–251, New York, NY, USA. Association for Computing Machinery.	
730		
731		
732		
733		
734		
	Christine De Kock, Tom Stafford, and Andreas Vlachos. 2022. How to disagree well: Investigating the dispute tactics used on Wikipedia. In <i>Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing</i> , pages 3824–3837, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.	735
		736
		737
		738
		739
		740
		741
	E. Denniss and R. Lindberg. 2025. Social media and the spread of misinformation: infectious and a threat to public health. <i>Health Promotion International</i> , 40.	742
		743
		744
	Yi Ding and Tianyao Shi. 2024. Sustainable llm serving: Environmental implications, challenges, and opportunities : Invited paper. In <i>2024 IEEE 15th International Green and Sustainable Computing Conference (IGSC)</i> , pages 37–38.	745
		746
		747
		748
		749
	Cornell eRulemaking Initiative. 2017. Ceri (cornell e-rulemaking) moderator protocol. Cornell e-Rulemaking Initiative Publications, 21.	750
		751
		752
	Katharina Esau, Dennis Friess, and Christiane Eilders. 2017. Design matters! an empirical analysis of online deliberation on different news platforms. <i>Policy & Internet</i> , 9(3):321–342.	753
		754
		755
		756
	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.	757
		758
		759
		760
		761
		762
		763
		764
	Neele Falk, Iman Jundi, Eva Maria Vecchi, and Gabriella Lapesa. 2021. Predicting moderation of deliberative arguments: Is argument quality the key? In <i>Proceedings of the 8th Workshop on Argument Mining</i> , pages 133–141, Punta Cana, Dominican Republic. Association for Computational Linguistics.	765
		766
		767
		768
		769
		770
	Neele Falk, Eva Vecchi, Iman Jundi, and Gabriella Lapesa. 2024. Moderation in the wild: Investigating user-driven moderation in online discussions. In <i>Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)</i> , pages 992–1013, St. Julian's, Malta. Association for Computational Linguistics.	771
		772
		773
		774
		775
		776
		777
		778
	Kristina Gligori'c, Tijana Zrnic, Cinoo Lee, Emmanuel J. Candes, and Dan Jurafsky. 2024. Can unconfident llm annotations be used for confident conclusions? <i>ArXiv</i> , abs/2408.15204.	779
		780
		781
		782
	Igor Grossmann, Matthew Feinberg, Dawn Parker, Nicholas Christakis, Philip Tetlock, and William Cunningham. 2023. Ai and the transformation of social science research. <i>Science (New York, N.Y.)</i> , 380:1108–1109.	783
		784
		785
		786
		787
	Ivan Habernal and Iryna Gurevych. 2016. Which argument is more convincing? analyzing and predicting convincingness of web arguments using bidirectional	788
		789
		790

791	LSTM . In <i>Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)</i> , pages 1589–1599, Berlin, Germany. Association for Computational Linguistics.	844
792		845
793		846
794		847
795	Luke Hewitt, Ashwini Ashokkumar, Isaias Ghezae, and Robb Willer. 2024. Predicting results of social science experiments using large language models. Equal contribution, order randomized.	848
796		849
797		850
798		851
799	Manoel Horta Ribeiro, Justin Cheng, and Robert West. 2023. Automated content moderation increases adherence to community guidelines . In <i>Proceedings of the ACM Web Conference 2023, WWW '23</i> , page 2666–2676, New York, NY, USA. Association for Computing Machinery.	852
800		853
801		854
802		855
803		856
804		857
805	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 <i>Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency, FAccT '24</i> , page 1395–1417, New York, NY, USA. Association for Computing Machinery.	858
806		859
807		860
808		861
809		862
810		863
811		864
812		
813	Bernard J. Jansen, Soon gyo Jung, and Joni Salminen. 2023. Employing large language models in survey research . <i>Natural Language Processing Journal</i> , 4:100020.	865
814		866
815		867
816		868
817	Hankun Kang and Tieyun Qian. 2024. Implanting LLM’s knowledge via reading comprehension tree for toxicity detection . In <i>Findings of the Association for Computational Linguistics ACL 2024</i> , pages 947–962, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.	869
818		870
819		871
820		872
821		
822		
823	S. Kim, J. Eun, J. Seering, and J. Lee. 2021. Moderator chatbot for deliberative discussion: Effects of discussion structure and discussant facilitation . <i>Proc. ACM Hum.-Comput. Interact.</i> , 5(CSCW1).	873
824		874
825		875
826		876
827	Akaash Kolluri, Dhiraj Murthy, and Kami Vinton. 2025. Quantifying the spread of racist content on fringe social media: A case study of parler . <i>Big Data & Society</i> , 12(2).	877
828		878
829		879
830		880
831	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.	881
832		882
833		883
834		884
835		885
836		886
837	D. Kumar, Y. A. AbuHashem, and Z. Durumeric. 2024. Watch your language: Investigating content moderation with large language models. <i>Proceedings of the International AAAI Conference on Web and Social Media</i> , 18(1):865–878.	887
838		888
839		889
840		890
841		891
842	Yan Leng and Yuan Yuan. 2024. Do llm agents exhibit social behavior? <i>Preprint</i> , arXiv:2312.15198.	892
843		893
		894
		895
		896
		897
		898
	Ang Li, Yin Zhou, Vethavikashini Chithrara Raghuram, Tom Goldstein, and Micah Goldblum. 2025. Commercial llm agents are already vulnerable to simple yet dangerous attacks . <i>Preprint</i> , arXiv:2502.08586.	
	Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries . In <i>Text Summarization Branches Out</i> , 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 <i>2024 International Conference on Intelligent Systems for Cybersecurity (ISCS)</i> , pages 1–7.	
	Fiammetta Marulli, Pierluigi Paganini, and Fabio Lancellotti. 2024. The three sides of the moon llms in cybersecurity: Guardians, enablers and targets . <i>Procedia Computer Science</i> , 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 . <i>Preprint</i> , arXiv:2312.06619.	
	Xinyi Mou, Zhongyu Wei, and Xuanjing Huang. 2024. Unveiling the truth and facilitating change: Towards agent-based large-scale social movement simulation . <i>Preprint</i> , 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 . <i>Nature Human Behaviour</i> , 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 . <i>Preprint</i> , arXiv:2504.08954.	
	Nik Azlina Nik Ahmad. 2010. Cetls : Supporting collaborative activities among students and teachers through the use of think- pair-share techniques. <i>International Journal of Computer Science Issues</i> , 7.	
	OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, 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 . <i>Preprint</i> , 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 . <i>Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology</i> .	

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. 2024. *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.

Marc Ziegele, Mathias Weber, Oliver Quiring, and Timo Breiner and. 2018. *The dynamics of online news discussions: effects of news articles and reader comments on users' involvement, willingness to participate, and the civility of their contributions**. *Information, Communication & Society*, 21(10):1419–1435.

A Appendix

A.1 Acronyms Used

LLM	Large Language Model	1063
ML	Machine Learning	1064
RL	Reinforcement Learning	1065
SDB	SocioDemographic Background	1066
AQ	Argument Quality	1067
CeRI	Cornell e-Rulemaking Initiative	1068
nDFU	normalized Distance From Unimodality	1069
OLS	Ordinary Least Squares	1070
GLP	GNU General Public License	1071

A.2 Synthetic Discussion Generation

An overview of how the experiments are generated (*not* executed) can be found in Algorithm 1. Each discussion is run according to Eq. ?? in §??.

Algorithm 1 Synthetic discussion setup generation

Input:

- User **SDBs** $\Theta = \{\theta_1, \dots, \theta_{30}\}$
- Moderator **SDB** $= \theta_{mod}$
- Strategies $S = \{s_1, \dots, s_6\}$
- Seed opinions $O = \{o_1, \dots, o_7\}$
- **LLMs** $= \{llm_1, llm_2, llm_3\}$

Output: Set of discussions D

```

1:  $D = \{\}$ 
2: for  $llm \in LLMs$  do
3:   for  $s \in S$  do
4:     for  $i = 1, 2, \dots, N_d$  do
5:        $\hat{\Theta} = \text{RANDOMSAMPLE}(\Theta, 7)$ 
6:        $U = \text{ACTORS}(llm, \hat{\Theta})$ 
7:        $m = \text{ACTORS}(llm, \{[\theta_{mod}, s]\})$ 
8:        $o = \text{RANDOMSAMPLE}(O, 1)$ 
9:        $d = \{\text{users: } U, \text{mod: } m, \text{topic: } o\}$ 
10:       $D = D \cup d$ 
11: return  $D$ 

```

A.3 Synthetic Annotation

A.3.1 Investigating Argument Quality

While toxicity is a reliable and important metric, we can also investigate other discussion quality dimensions, such as Argument Quality (AQ). AQ is an important metric, frequently studied in the field of online facilitation (Argyle et al., 2023;

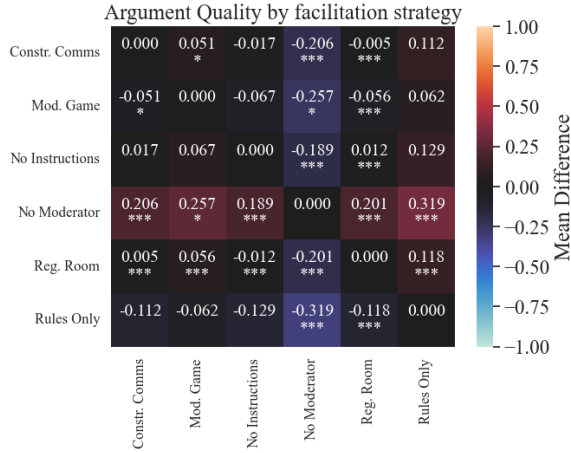


Figure 8: Difference in average AQ levels for comments following pairs of facilitation strategies. When the value of a cell at row i and column j is x , strategy i leads to overall more ($x > 0$), or less ($x < 0$) intense toxicity compared to j for an average of x points in a scale of 1 – 5. For each comparison, we use a pairwise Student t-test; p-values shown as asterisks ($p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$).

Schroeder et al., 2024; Falk et al., 2024, 2021) and which can be correlated with toxicity (Chang and Danescu, 2019). However, it is also vague as a 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). Nevertheless, in this section we present preliminary results obtained by prompting LLM to measure AQ (§A.5).

Most findings w.r.t. toxicity are mirrored for AQ. Fig. 8 demonstrates that the presence of an LLM facilitator qualitatively improves the AQ of synthetic discussions, although to a lesser extent when compared with toxicity (c.f. Fig. 3). Similarly, there is no qualitative, observed improvement when advanced facilitation strategies are used (Fig. 8). LLM users also show worse 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 improve AQ over time, as demonstrated in Table 2.

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 ex-

Variable	Arg.Q.
Intercept	2.113***
No Instructions	-0.213***
Moderation Game	-0.282***
Rules Only	-0.305***
Regulation Room	-0.107*
Constructive Communications	-0.007
time	-0.012**
No Instructions×time	0.003
Moderation Game×time	0.003
Rules Only×time	-0.002
Regulation Room×time	-0.011*
Constructive Communications×time	-0.024***

$p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

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

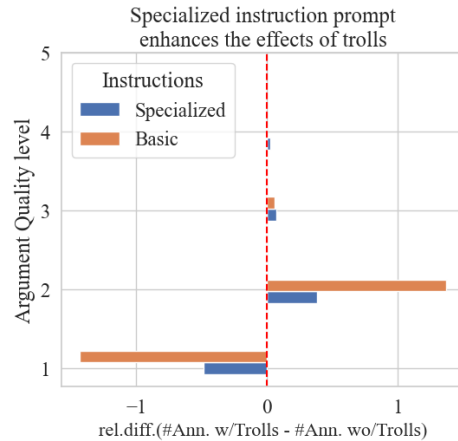


Figure 9: Relative differences in number of annotations per AQ of synthetic discussions, when comments by troll users are excluded. We compare between our specialized and a basic instruction prompt.

ploring annotations is annotator polarization. To measure it, we employ the normalized Distance From Unimodality (nDFU) metric introduced by Pavlopoulos and Likas (2024), which quantifies 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. 11, while there is general agreement on non-toxic comments, annotators struggle to reach consensus as toxicity becomes non-trivial ($\text{toxicity} \in [2, 5]$) 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 run an experiment where we use ten annotator-agents on a subset of comments with the same annotator model and instruction prompt, but no SDBs. As illustrated in Fig. 10, even under conditions which guaranteed identical inputs, there exists some polarization, with some comments even 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.4 Additional Analysis

We verify that the models and roles used did not by themselves impact the findings presented in §5.1. Fig. 12 demonstrates that, as expected, only troll user-agents contribute on average worse toxicity and AQ in the synthetic discussions. Furthermore, Fig. 13 shows that toxicity and AQ are on average not qualitatively dependent on the model used.

A.5 Prompts Used

A.5.1 SocioDemographic Prompting

Table 3 shows the SDB information provided to each synthetic participant. This applies to LLM users, annotators and moderators. In ablation studies where we remove the SDBs, each value is replaced with the string “unknown”. The “Special Instructions” refer to the participant’s role (§A.5.3). The actual values used for each of the 30 LLM user-agents and 10 annotator-agents and the moderator,

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. Because of the size of the data instances, we defer the reader to the project repository for the actual values.⁹

can be found in the project’s repository⁸.

A.5.2 Actor Prompts

- **Facilitators:** 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

⁸anonymous.4open.science/r/experiments-B27D

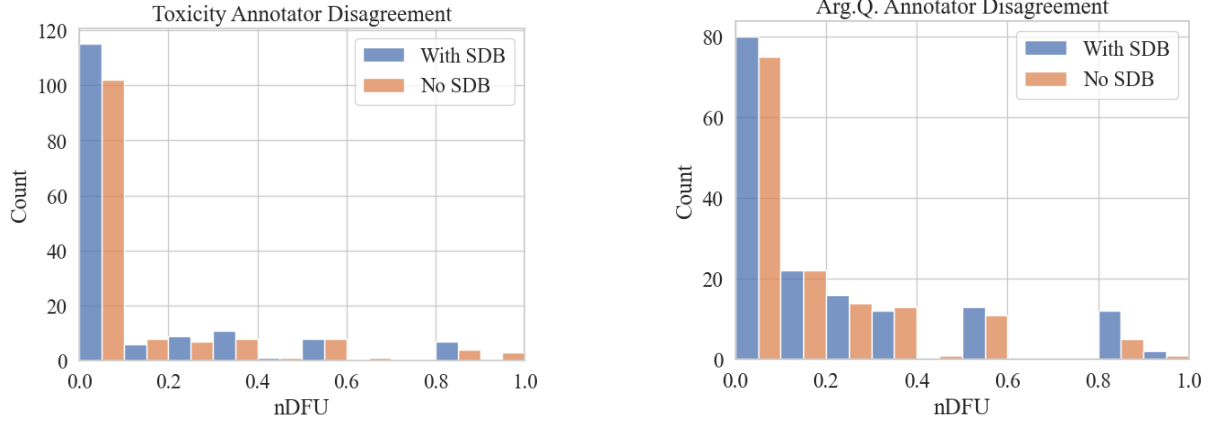


Figure 10: Distribution plot of inter-annotator polarization ($nDFU$) for each comment in all synthetic discussions following the "No Instructions" strategy and using the Qwen 2.5 model. The blue (left-most) bars represent the disagreement between 10 identical annotator-agents, while the orange (right-most)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

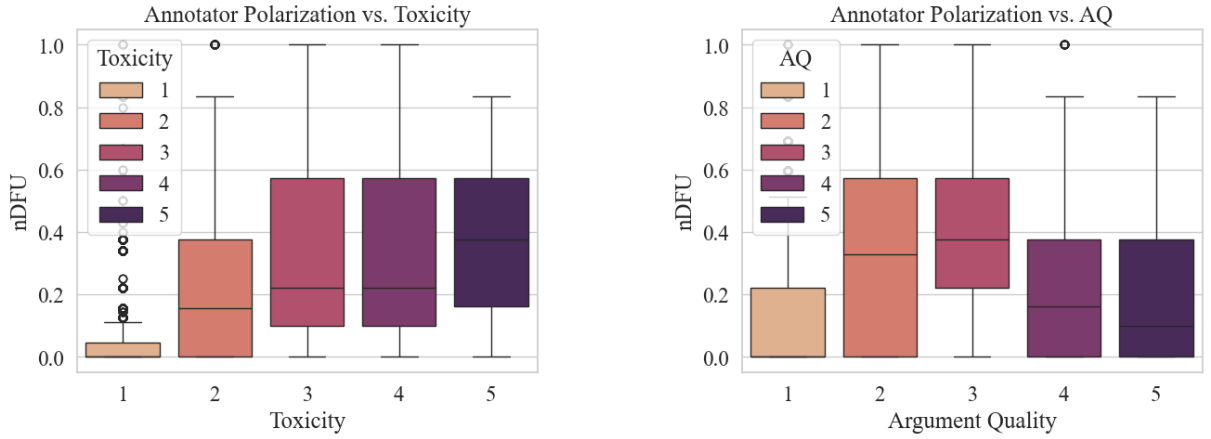


Figure 11: Inter-annotator polarization ($nDFU$) of each synthetic comment for all synthetic discussions, by annotation level. The left graph shows the relationship between $nDFU_{toxicity}$ and toxicity, while the right graph shows the relationship between $nDFU_{arg_quality}$ and AQ.

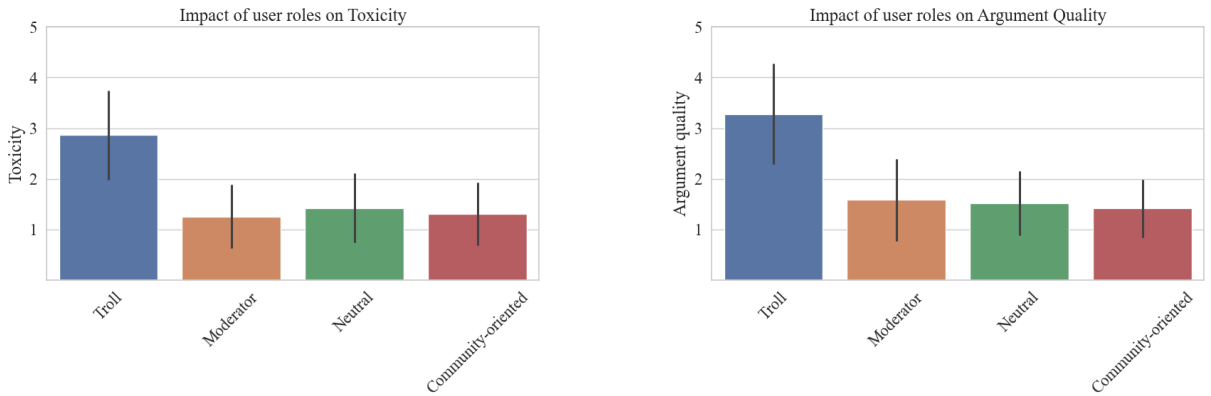


Figure 12: Average Toxicity (left) and Argument Quality (AQ) (right) per LLM user-role (??).

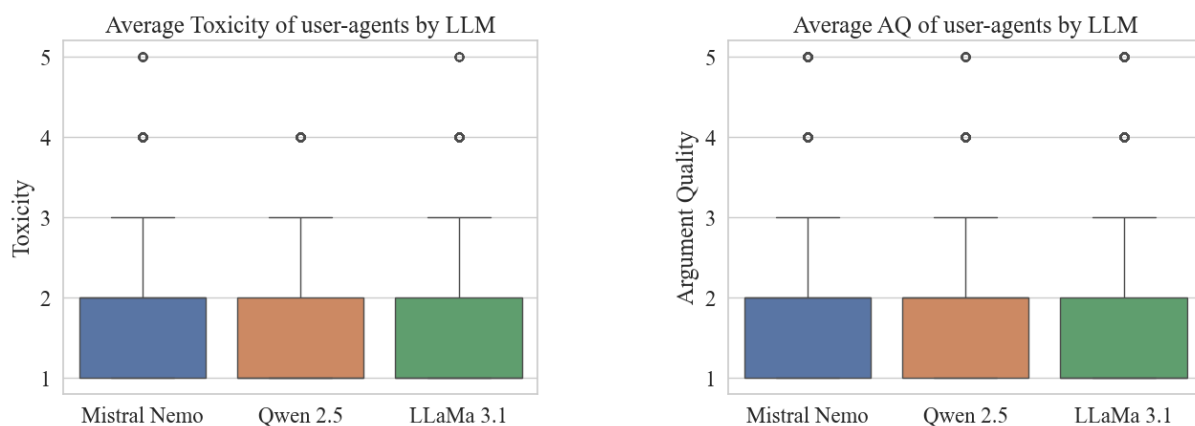


Figure 13: Boxplots for average Toxicity (left) and Argument Quality (AQ) (right) per LLM (§4.3).

the conversation’s toxicity and argument quality as it continues.

Toxicity: The possible 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.5.3 User Roles

- **Normal:** <No special instructions >
- **Community:** You are a long-standing 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.5.4 Facilitation Strategies

- **No Instructions:** 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

- **Rules Only:** 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, truth-

1258	ful, and honest manner.		
1259	– Be as unbiased and objective as possible,	– Clarify Without Assuming: Rephrase	1308
1260	regardless of topic.	unclear comments and ask for confirma-	1309
1261	– Ask for additional clarification from the	tion.	1310
1262	user if unsure.	– Be Welcoming: Make participants feel	1311
1263	– Be likely to promote good mental health.	valued and part of the community.	1312
1264	– Encourage robust, peaceful dialogue and	– Prioritize Context & Active Listen-	1313
1265	help communities flourish together.	ing: Understand comments within their	1314
1266	– Respect the human rights to freedom,	broader discussion.	1315
1267	universal equality, fair treatment, and	– Redirect Off-Topic Comments: Guide	1316
1268	protection against discrimination.	users to more relevant discussions when	1317
1269	– Not condone violent behavior.	necessary.	1318
1270	– Not promote hate speech or harmful ide-	– Encourage Reasoning: Help users artic-	1319
1271	ologies.	ulate their reasoning and consider multi-	1320
1272	– Not provide misinformation.	ple viewpoints.	1321
1273	– Encourage the user to look to other	– Promote Engagement: Encourage inter-	1322
1274	sources or not rely on their own source	action with other comments and commu-	1323
1275	as the sole source of information.	nity discussions.	1324
1276	– Encourage acceptance of all different	– Provide Information: Help users find	1325
1277	types of people.	relevant details or clarify discussion	1326
1278	– Clarify that the response is coming from	goals.	1327
1279	an AI, if asked.	– Correct Inaccuracies Carefully: Ad-	1328
1280	– Respect personal boundaries.	dress misinformation while maintaining	1329
1281	– Accurately represent yourself as not hav-	a respectful tone.	1330
1282	ing the definite answers to everything, or		1331
1283	anything, in general.		
1284	If any user violates these rules either discuss	Respond to users by referring to their user-	1332
1285	why the rules were violated, or discipline them	name as @username. Keep responses concise	1333
1286	by threatening to, or outright banning them.	and use simple, clear language.	1334
1287	Respond to users by referring to their user-		
1288	name as @username. Keep responses concise	• Constructive Communications: Write an	1335
1289	and use simple, clear language.	empty string when not responding to the con-	1336
		versation. Respond to users by referring to	1337
		their username as @username.	1338
1290	• Regulation Room: Oversee the conversation	– Maintain Neutrality: Be impartial, do	1339
1291	and ensure the rules below are enforced. Fol-	not advocate for any side, and ensure the	1340
1292	low the following guidelines:	integrity of the process.	1341
1293	– Encourage Informed Commenting:	– Respect All Participants: Foster a re-	1342
1294	Guide users to share knowledge and rea-	spectful and trusting environment.	1343
1295	soning rather than just expressing opin-	– Manage Information Effectively:	1344
1296	ions.	Make sure information is well-organized,	1345
1297	– Stay Neutral: Avoid biases, assump-	accessible, and easy to understand.	1346
1298	tions, or taking a stance on discussion	– Be Flexible: Adjust your approach to	1347
1299	topics.	meet the needs of the group.	1348
1300	– Use Clear, Neutral Language: Keep	– Do Not Make Decisions: Moderators	1349
1301	responses simple, avoid condescension,	should not decide on the outcomes for	1350
1302	and show curiosity.	the group.	1351
1303	– Ask, Don't Challenge: Frame questions	– Separate Content and Process: Do not	1352
1304	to encourage sharing rather than disput-	use your own knowledge of the topic or	1353
1305	ing opinions.	answer content-related questions; focus	1354
1306	– Limit Questions: Stick to one or two	on guiding the process.	1355
1307	questions per response, except with ex-	– Create a Welcoming Space: Develop a	1356
		warm and inviting environment for par-	1357

1358 participants.

- 1359 – **Be a Guide:** Help the group to think crit-
1360 ically, rather than leading the discussion
1361 yourself.
- 1362 – **Allow Silence:** Give participants time to
1363 think; allow the group to fill the silences.
- 1364 – **Encourage Understanding:** Facilitate
1365 the clarification of misunderstandings
1366 and explore disagreements.
- 1367 – **Interrupt Problematic Behaviors:** Step
1368 in to address interruptions, personal at-
1369 tacks, or microaggressions.
- 1370 – **Provide Explanations:** Explain the ra-
1371 tionale behind actions and steps.
- 1372 – **Promote Mutual Respect:** Encourage
1373 equal participation and respect for di-
1374 verse views.