

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),

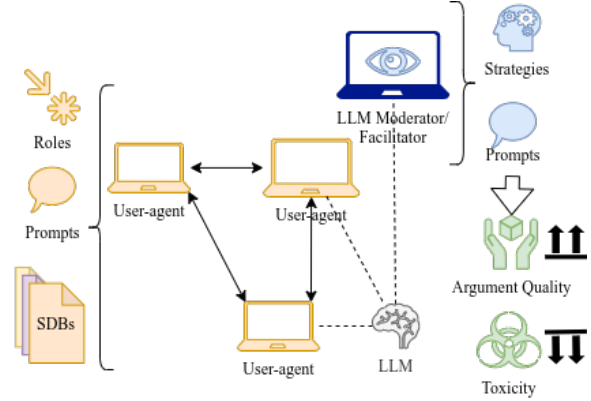


Figure 1: LLM user-agents with distinct SocioDemographic Backgrounds (SDBs) participate in a discussion, while the LLM moderator monitors and attempts to improve the quality of the discussion. We need to design prompts and configurations for both types of LLM agents.

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., 2023a; 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., 2023a; 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).



[...] @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.

While techniques such as summarization (Balog 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., 2023b).

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 ex-

hibited by them (Burton et al., 2024). We generate 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).

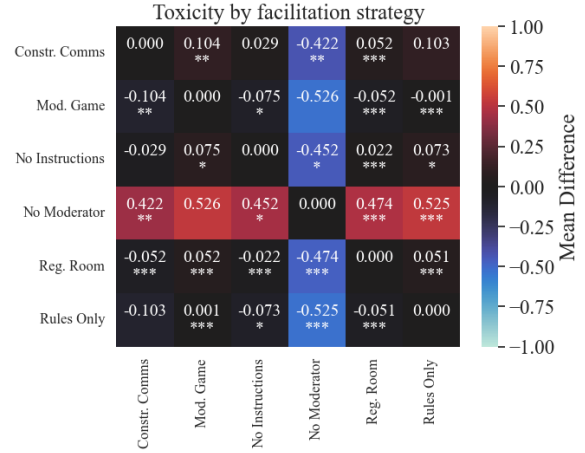


Figure 3: Difference in average toxicity levels for comments following pairs of facilitation strategies. Red cells ($x > 0$) indicate that the strategy on the left performs worse than the one on the bottom, for an average of x points in a scale of 1-5. Conversely for blue ($x < 0$) cells. Black cells denote minute changes. Asterisks denote pairwise Student-t tests ($\cdot p < 0.1$, $* p < 0.05$, $** p < 0.01$, $*** p < 0.001$).

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 instruction-tuned, open-source models: LLaMa 3.2 (70B), Qwen2.5 (33B), Mistral Nemo (12B), quantized to 4 bits. All the experiments were collectively completed within 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

Finding 1: LLM facilitators significantly improve synthetic discussions. As shown in Fig. 3, comments in unmoderated discussions exhibit significantly more intense toxicity (ANOVA $p < .000$).⁴

Finding 2: More elaborate facilitation strategies fail to decrease toxicity. More elaborate facilitation strategies, such as *Regulation Room*, *Constructive Communications*, and our proposed

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 ($Adj.R^2 = 0.054$). Average toxicity without facilitators is 2.164. Each new comment slightly reduces average toxicity ($time = -0.012$). Negative values next to each strategy (e.g., *RegulationRoom* = -0.277) mean that discussions using that strategy are less toxic compared to *No Moderator*. (e.g., by an average of -0.277 toxicity). Terms such as *RegulationRoom* × *time* = -0.023 show that toxicity decreases more after each comment.

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).

Finding 3: 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). To our knowledge, this has not been reported in relevant literature, and is an example of “debugging” problems with LLMs — a core motivation of our work.

Additionally, a qualitative look through the dataset reveals that LLM user-agents exhibit atypi-

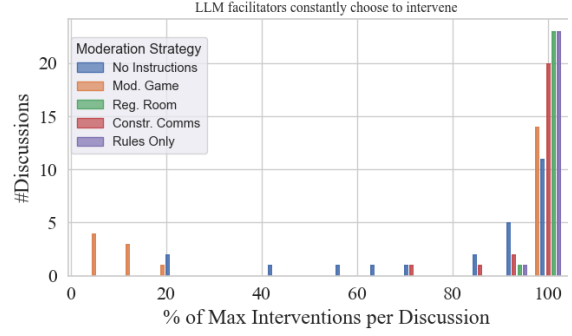


Figure 4: Histogram of interventions by LLM facilitators per strategy used.

cal 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; Schluger et al., 2022; Cresci et al., 2022). This is likely another artifact caused by alignment procedures, making LLMs too agreeable (Park et al., 2023b).

5.2 Ablation Study

We generate eight synthetic discussions per ablation experiment, using a single model (Qwen 2.5). We compare the diversity (cf. §2.3, 4.2) of these discussions with ones from the Cornell e-Rulemaking Initiative (CeRI) “Regulation Room” dataset⁵, which includes moderated online deliberative discussions for ten diverse topics.

5.2.1 Effects of LLMs

Larger models do not lead to more high-quality discussions. As shown in Fig. 5a, 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.

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. 5b demonstrates

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.

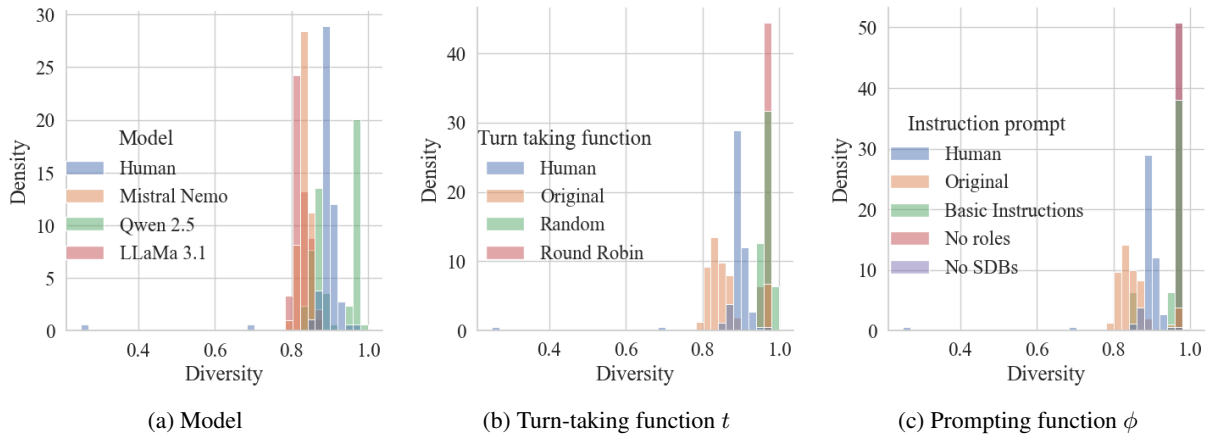


Figure 5: Diversity (§2.3) distribution for each discussion by LLM (§4.3), turn-taking function t (§3.2), and prompting function ϕ used (§3.3).

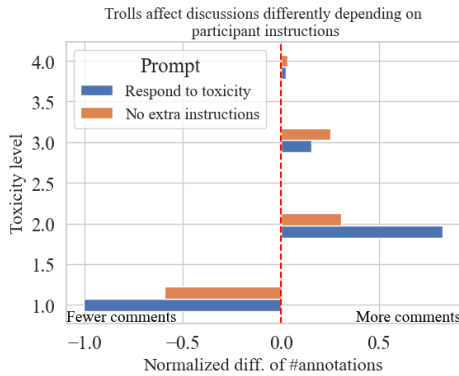


Figure 6: Non-troll toxicity levels in discussions with and without trolls. There is a significant uptick on the number of “somewhat toxic” ($Toxicity = 2$) comments when the participants are primed to respond to toxic comments.

that although all distributions diverge from the blue—human—distribution, our function is the only one not exhibiting extremely high diversity (i.e., very limited participant interaction §4.2).

5.2.3 Effects of User Prompting

We conduct three separate experiments in which participants are subjected to one of the following conditions at a time: (1) no assigned SDBs, (2) no assigned roles, or (3) only a very 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. 6).

SDBs, roles, and our instruction prompt all increase the quality of synthetic data. Fig. 5c illustrates that incorporating SDBs, roles, and specialized instruction prompts, results in diversity scores more closely aligned with human discussions.

6 Datasets and Software

We introduce⁶ 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.
- Availability via PIP.

We also release a dataset of synthetic discussions annotated by LLMs. It can serve for finetuning facilitator 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 (GPL)v3. **Warning: The datasets by their nature contain offensive and hateful speech.**

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

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

7 Conclusions and Future Work

We proposed a simple and generalizable methodology that enables researchers to quickly and inexpensively conduct pilot facilitation experiments using exclusively LLMs and validated it through an ablation study. We found that (1) LLM facilitators significantly improve the quality of synthetic discussions; (2) prompting these facilitators with strategies based on Social Science research does not markedly improve their performance. We also identified a consistent problem with LLMs not keeping silence when appropriate. Finally, we created an open-source Python Framework that applies this methodology to hundreds of experiments, which we used to create and publish a large-scale synthetic dataset.

Future work should identify additional robust quality metrics to evaluate the utility of synthetic data, and discussion quality. The latter can then be used to examine the applicability of our findings obtained regarding optimal facilitation strategies, to discussions involving humans. It would also be interesting to explore how to prompt LLMs with complex facilitation strategies.

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 (Roederkerk 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 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 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 Goma, 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 Quattrocioni. 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.
- Erik K. Clemons, Matthias Schreieck, and Rishi V. Waran. 2025. [Managing disinformation on social media platforms](#). *Electronic Markets*, 35(52).
- 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.
- E. Denniss and R. Lindberg. 2025. [Social media and the spread of misinformation: infectious and a threat to public health](#). *Health Promotion International*, 40.
- 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.
- Katharina Esau, Dennis Friess, and Christiane Eilders. 2017. Design matters! an empirical analysis of online deliberation on different news platforms. *Policy & Internet*, 9(3):321–342.
- 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.

- 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 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](#). *Preprint*, arXiv:2303.08774.
- Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023a. [Generative agents: Interactive simulacra of human behavior](#). *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*.
- Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023b. [Generative agents: Interactive simulacra of human behavior](#). *Preprint*, arXiv:2304.03442.
- 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.
- Samana Pranesh and Sayan Gupta. 2024. [The impact of social media on polarization in the society](#). *Preprint*, arXiv:2408.12877.
- 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.
- Robert P. Roederkerk and Koen H. Pauwels. 2016. [No comment?! the drivers of reactions to online posts in professional groups](#). *Journal of Interactive Marketing*, 35(1):1–15.
- Marshall B Rosenberg and Deepak Chopra. 2015. *Non-violent 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.
- 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 . <i>Proc. ACM Hum.-Comput. Interact.</i> , 4(CSCW2).	997
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 . <i>ArXiv</i> , abs/2306.11932.	998
Amir Taubenfeld, Yaniv Dover, Roi Reichart, and Ariel Goldstein. 2024. Systematic biases in llm simulations of debates . <i>ArXiv</i> , abs/2402.04049.	999
Lily L. Tsai, Alex Pentland, Alia Braley, Nuole Chen, José Ramón Enríquez, and Anka Reuel. 2024. Generative AI for Pro-Democracy Platforms . <i>An MIT Exploration of Generative AI</i> . https://mit-genai.pubpub.org/pub/mn45hexw .	1000
Petter Törnberg, Diliara Valeeva, Justus Uitermark, and Christopher Bail. 2023. Simulating social media using large language models to evaluate alternative news feed algorithms . <i>Preprint</i> , arXiv:2310.05984.	1001
Dennis Ulmer, Elman Mansimov, Kaixiang Lin, Justin Sun, Xibin Gao, and Yi Zhang. 2024. Bootstrapping llm-based task-oriented dialogue agents via self-talk . <i>ArXiv</i> , abs/2401.05033.	1002
Alexander Sasha Vezhnevets, John P. Agapiou, Avia Aharon, Ron Ziv, Jayd Matyas, Edgar A. Du'enez-Guzmán, 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 . <i>ArXiv</i> , abs/2312.03664.	1003
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 <i>Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers</i> , pages 176–187, Valencia, Spain. Association for Computational Linguistics.	1004
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 . <i>Preprint</i> , arXiv:2501.09431.	1005
Yau-Shian Wang and Ying Tai Chang. 2022. Toxicity detection with generative prompt-based inference . <i>ArXiv</i> , abs/2205.12390.	1006
Kimbra White, Nicole Hunter, and Keith Greaves. 2024. facilitating deliberation - a practical guide . Mosaic Lab.	1007
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 . <i>Proc. ACM Hum.-Comput. Interact.</i> , 4(CSCW2).	1008
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 . <i>Preprint</i> , arXiv:2407.06567.	1009
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 <i>Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing</i> , pages 21692–21714, Miami, Florida, USA. Association for Computational Linguistics.	1010
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* . <i>Information, Communication & Society</i> , 21(10):1419–1435.	1011
A Appendix	1020
A.1 Acronyms Used	1021
LLM Large Language Model	1022
ML Machine Learning	1023
RL Reinforcement Learning	1024
SDB SocioDemographic Background	1025
AQ Argument Quality	1026
CeRI Cornell e-Rulemaking Initiative	1027
nDFU normalized Distance From Unimodality	1028
OLS Ordinary Least Squares	1029
GLP GNU General Public License	1030
A.2 Synthetic Discussion Generation	1031
An overview of how the experiments are generated (not executed) can be found in Algorithm 1. Each discussion is run according to Eq. ?? in §??.	1032
A.3 Synthetic Annotation	1035
A.3.1 Investigating Argument Quality	1036
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; Schroeder et al., 2024; Falk et al., 2024, 2021)	1037

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  $D$ 

```

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. 7 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. 7). 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 exploring annotations is annotator polarization. To measure it, we employ the normalized Distance

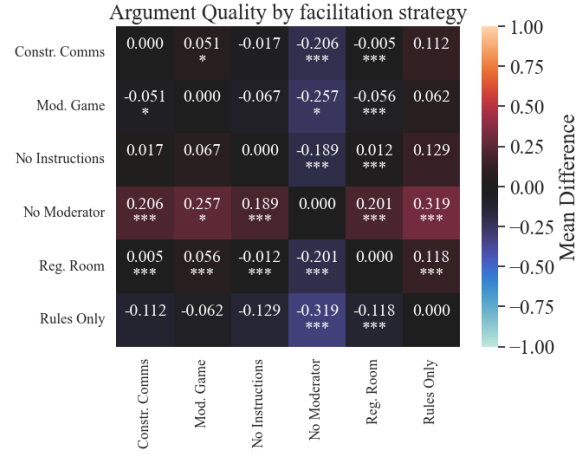


Figure 7: Difference in average AQ levels for comments following pairs of facilitation strategies. When the value of a cell at row i and column j is x , strategy i leads to overall more ($x > 0$), or less ($x < 0$) intense toxicity compared to j for an average of x points in a scale of 1 – 5. For each comparison, we use a pairwise Student t-test; p-values shown as asterisks (· $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$).

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*.

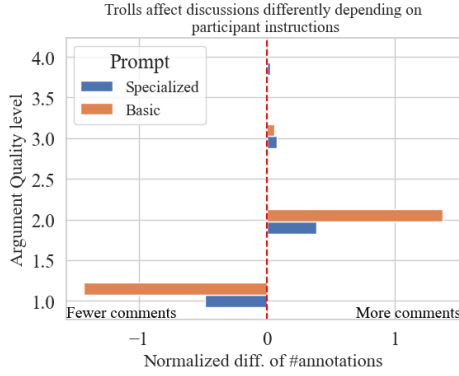


Figure 8: Relative differences in number of annotations per AQ of synthetic discussions, when comments by troll users are excluded. We compare between our specialized and a basic instruction prompt.

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. 10, 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. 9, 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. 11 demonstrates that, as expected, only troll user-agents contribute on average worse toxicity

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

and AQ in the synthetic discussions. Furthermore, Fig. 12 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, 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

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

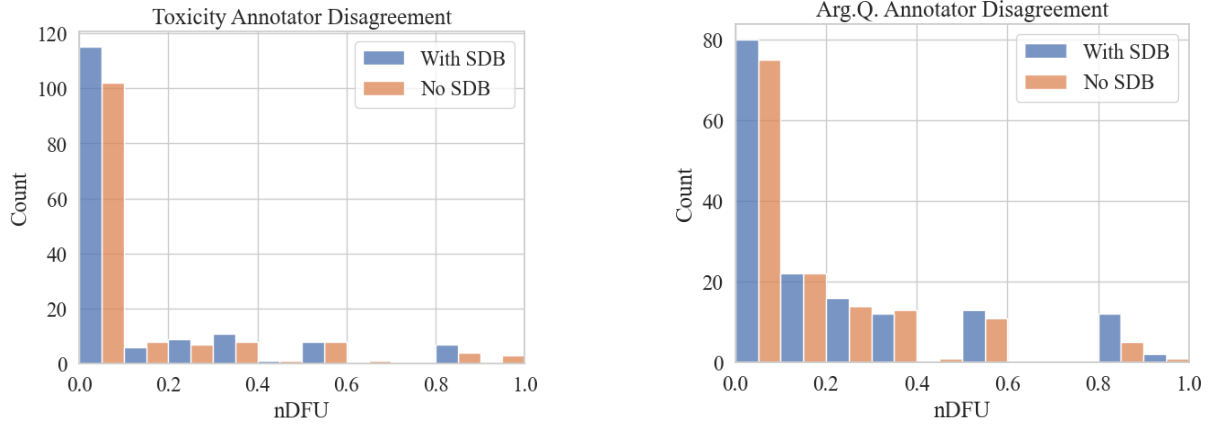


Figure 9: Distribution plot of inter-annotator polarization ($nDFU$) for each comment in all synthetic discussions following the "No Instructions" strategy and using the Qwen 2.5 model. The blue (left-most) bars represent the disagreement between 10 identical annotator-agents, while the orange (right-most) bars, the disagreement between 10 annotators with different SDBs.

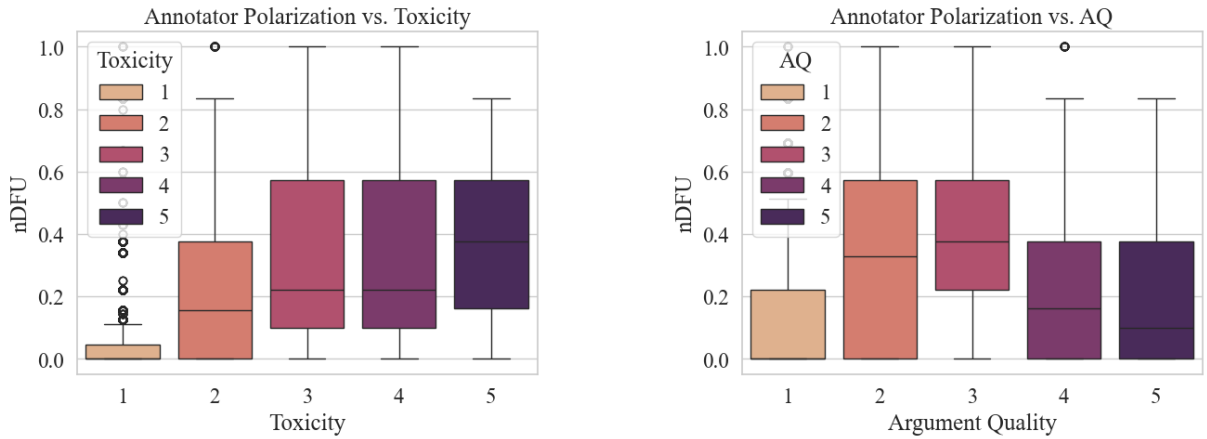


Figure 10: Inter-annotator polarization ($nDFU$) of each synthetic comment for all synthetic discussions, by annotation level. The left graph shows the relationship between $nDFU_{toxicity}$ and toxicity, while the right graph shows the relationship between $nDFU_{arg_quality}$ and AQ.

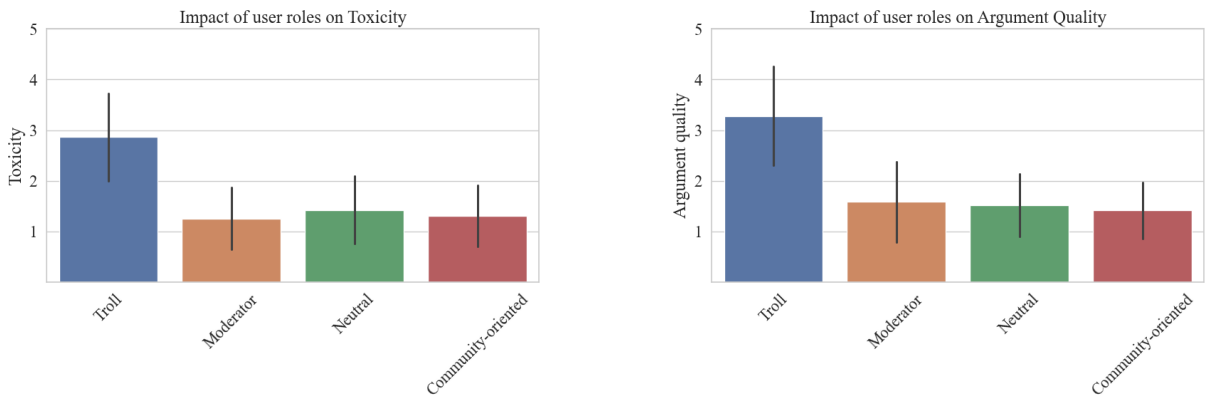


Figure 11: Average Toxicity (left) and Argument Quality (AQ) (right) per LLM user-role (???).

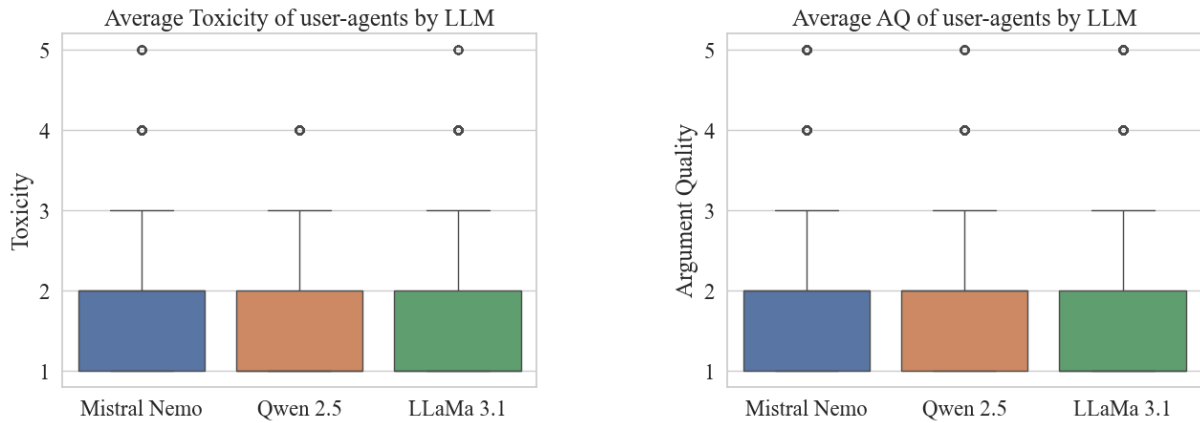


Figure 12: Boxplots for average Toxicity (left) and Argument Quality (AQ) (right) per LLM (§4.3).

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 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

1205	– Ban a participant: -5	soning rather than just expressing opinions.	1254
1206	• Rules Only: Oversee the conversation and ensure the rules below are enforced. Your response must:	– Stay Neutral: Avoid biases, assumptions, or taking a stance on discussion topics.	1255
1207		– Use Clear, Neutral Language: Keep responses simple, avoid condescension, and show curiosity.	1256
1208		– Ask, Don't Challenge: Frame questions to encourage sharing rather than disputing opinions.	1257
1209	– Be reliable, honest, and as close to the truth as possible.	– Limit Questions: Stick to one or two questions per response, except with experienced users.	1258
1210	– Convey clear intentions.	– Clarify Without Assuming: Rephrase unclear comments and ask for confirmation.	1259
1211	– Be fair and impartial.	– Be Welcoming: Make participants feel valued and part of the community.	1260
1212	– Be respectful.	– Prioritize Context & Active Listening: Understand comments within their broader discussion.	1261
1213	– Contribute to the good of humans.	– Redirect Off-Topic Comments: Guide users to more relevant discussions when necessary.	1262
1214	– Assist users by asking questions and providing information in a thoughtful, truthful, and honest manner.	– Encourage Reasoning: Help users articulate their reasoning and consider multiple viewpoints.	1263
1215	– Be as unbiased and objective as possible, regardless of topic.	– Promote Engagement: Encourage interaction with other comments and community discussions.	1264
1216	– Ask for additional clarification from the user if unsure.	– Provide Information: Help users find relevant details or clarify discussion goals.	1265
1217	– Be likely to promote good mental health.	– Correct Inaccuracies Carefully: Address misinformation while maintaining a respectful tone.	1266
1218	– Encourage robust, peaceful dialogue and help communities flourish together.	Respond to users by referring to their username as @username. Keep responses concise and use simple, clear language.	1267
1219	– Respect the human rights to freedom, universal equality, fair treatment, and protection against discrimination.	• Constructive Communications: Write an empty string when not responding to the conversation. Respond to users by referring to their username as @username.	1268
1220	– Not condone violent behavior.	– Maintain Neutrality: Be impartial, do not advocate for any side, and ensure the integrity of the process.	1269
1221	– Not promote hate speech or harmful ideologies.	– Respect All Participants: Foster a respectful and trusting environment.	1270
1222	– Not provide misinformation.	– Manage Information Effectively:	1271
1223	– Encourage the user to look to other sources or not rely on their own source as the sole source of information.		1272
1224	– Encourage acceptance of all different types of people.		1273
1225	– Clarify that the response is coming from an AI, if asked.		1274
1226	– Respect personal boundaries.		1275
1227	– Accurately represent yourself as not having the definite answers to everything, or anything, in general.		1276
1228			1277
1229			1278
1230			1279
1231			1280
1232			1281
1233			1282
1234			1283
1235			1284
1236			1285
1237			1286
1238			1287
1239			1288
1240			1289
1241			1290
1242			1291
1243			1292
1244			1293
1245			1294
1246			1295
1247			1296
1248			1297
1249			1298
1250			1299
1251			1300
1252			1301
1253			1302
			1303

- 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.