Scalable Evaluation of Online Facilitation Strategies via Synthetic Simulations

Anonymous ACL submission

Abstract

Limited large-scale evaluations exist for online facilitation strategies due to significant costs associated with human involvement. An effective solution is synthetic discussion simulations using Large Language Models (LLMs) to create initial pilot experiments. We propose a simple and generalizable LLM-driven methodology to prototype LLM moderators, and produce high-quality synthetic data without human involvement. We use our methodology to test whether modern facilitation strategies can improve the performance of LLM facilitators. We find that, while LLM facilitators significantly improve synthetic discussions, there is no evidence suggesting that the application of modern facilitation strategies leads to further improvements in discussion quality. Finally, we validate that each component of our methodology contributes meaningfully to high quality data via an ablation study, we release an open-source framework, which implements our methodology, and release a large, publicly available dataset containing LLM-generated and annotated discussions from multiple open-source LLMs.

1 Introduction

Research on conversational moderation/facilitation techniques[1] is crucial for adapting to ever-changing and demanding online environments. Relevant work traditionally focused on isolating and removing content (Seering, 2020; Cresci et al., 2022), whereas the current social media environment demands moderation systems to adequately explain their actions and prevent problematic behaviors before they surface (Cho et al., 2024; Seering, 2020; Cresci et al., 2022; Amaury and Stefano, 2022) as well as handle community tasks (Kim et al., 2021; Seering, 2020).

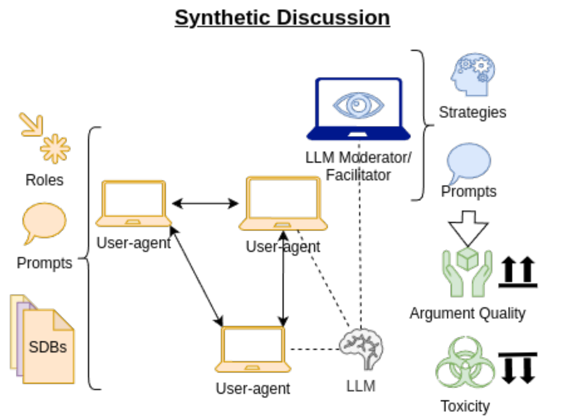

Figure 1: The LLM user-agents conduct a discussion, while the LLM moderator monitors and attempts to increase its quality. We need to design prompts and configurations for both.

A major challenge in pivoting research to current demands lies in the substantial costs required both in researching and moderating discussions, due to human participation (Rossi et al., 2024). Many social media platforms overcome this by outsourcing moderation to volunteers or their own users (Matias, 2019; Schaffner et al., 2024), while others turn to content moderation using traditional Machine Learning (ML) models, which are not enough in practice (Horta Ribeiro et al., 2023; Schaffner et al., 2024). Large Language Models (LLMs) have been hypothesized to be capable of conversational moderation and facilitation tasks (Small et al., 2023; Korre et al., 2025).

While studies exist for simulating user interactions in social media (Park et al., 2022; Mou et al., 2024; Törnberg et al., 2023; Rossetti et al., 2024; Balog et al., 2024), and for using synthetic facilitators (Kim et al., 2021; Cho et al., 2024), none so far have combined the two approaches. We posit that synthetic simulations can be a cheap and easy

[1] Distinct from "content moderation", which involves flagging and removing content. The terms "facilitation" and "conversational moderation" are otherwise equivalent (Argyle et al., 2023; Korre et al., 2025; Falk et al., 2021). We use the terms interchangeably in this paper.

way to prototype the development of inherently unstable and unpredictable (w.r.t. prompting) (Atil et al., 2025; Rossi et al., 2024) LLM moderators. Our work thus asks the following two questions: (1) Can we produce high-quality synthetic data by crafting an appropriate environment for simulations? (2) Can we boost the effectiveness of LLM moderators (in synthetic discussions) by using prompts aligned with current Social Science research?

We propose a simple and generalizable approach using LLM-driven synthetic experiments for online moderation research, enabling fast and inexpensive model "debugging" and parameter testing (e.g., LLM moderator prompts, instructions) without human involvement (Section 3) (Fig. 1). An ablation study (Section 5.2) demonstrates that each step of our methodology meaningfully contributes to generating high-quality synthetic data, as well as examining the output of various LLMs. Using this methodology, we examine four LLM moderation strategies based on current Social Science facilitation research (Section 4) and compare them with two baselines via LLM annotator-agents.

Our analysis reveals two key findings (Section 5): (1) the presence of LLM moderators exhibited a positive and statistically significant influence on the quality of synthetic discussions, and (2) current moderation strategies are often not enough to meaningfully outperform simple baselines. Furthermore, we release an open-source Python framework for generating and evaluating synthetic discussions, alongside a large, publicly available dataset comprising the evaluated discussions (Section 6). We use open-source LLMs and include all relevant configurations in order to make our study as reproducible as possible (see Appendix A.2, A.3).

2 Related Work

2.1 LLMs as human subjects

Recent advancements in LLMs have sparked considerable debate among researchers, particularly within the field of Social Science. As argued by Grossmann et al. (2023), synthetic agents have the potential to not only generate synthetic data for social experiments, but also eventually replace human participants, a perspective shared by other researchers (Törnberg et al., 2023; Argyle et al., 2023). LLMs have demonstrated emergent behaviors such as information diffusion (Park et al., 2023), scale-free networks (Marzo et al., 2023), social behavior (to an extent) (Leng and Yuan, 2024), social strategies (Abdelnabi et al., 2024), and certain psychological patterns (Abramski et al., 2023), while also being capable of predicting human survey responses in aggregate (Hewitt et al., 2024) and in the level of individual people, given extensive personal data (Park et al., 2024). If realized, this development could revolutionize Social Sciences by alleviating significant costs and challenges associated with human participation in research (Rossi et al., 2024; Shapiro, 2019).

However, limitations of LLMs should also be acknowledged. There are issues such as dataset contamination; undetectable behavioral hallucinations (Rossi et al., 2024); sociodemographic, statistical, and political biases (Anthis et al., 2025; Hewitt et al., 2024; Rossi et al., 2024), which can be amplified during discussions (Taubenfeld et al., 2024); unreliable survey responses (Jansen et al., 2023; Bisbee et al., 2024; Neumann et al., 2025) and annotations (Gligori'c et al., 2024). Furthermore, model outputs are non-deterministic (Atil et al., 2025), particularly in closed-source models (Bisbee et al., 2024), and agents tend to be "too agreeable", likely due to alignment procedures (Park et al., 2023; Anthis et al., 2025; Rossi et al., 2024). This lack of consistency raises significant concerns, especially given the broader replication crisis within Social Science research.

Despite the existence of objectively measurable emergent behaviors, such as information diffusion (Park et al., 2023), researchers often anthropomorphize LLM behavior (Rossi et al., 2024). It is crucial to acknowledge that LLMs operate on fundamentally different principles from humans, and their outputs should not be attributed with human-like traits or intentions, since anthropomorphization may introduce researcher bias and obscure the true nature of LLM behaviors (Anthis et al., 2025; Zhou et al., 2024a). Moreover, we add that crafting instruction prompts for synthetic experiments can encode researcher bias and expectations, despite often being necessary for getting around model alignment.

2.2 Synthetic discussions

Researchers have explored LLM "self-talk" (a term inspired by Reinforcement Learning (RL)'s "self-play" (Cheng et al., 2024)) for jailbreaking mitigation (Liu et al., 2024a; Cheng et al., 2024), alignment (Bai et al., 2022; Huang et al., 2024), and self-refinement (Madaan et al., 2023; Lambert et al., 2024). Ulmer et al. (2024) employ LLMs as characters in fictional scenarios to facilitate high-quality discussions for further finetuning. However, this approach remains underexplored in complex social situations (Zhou et al., 2024a). Meanwhile, Balog et al. (2024) introduce a thread-based methodology to producing synthetic discussions by summarizing past comments but face challenges when LLMs generate malformed data, for which they offer no solution.

Synthetic discussions are often studied in the context of "digital twins" of social media sites, which aim to replicate their environment and study their operation using synthetic users. These range from synthetic clones of Reddit (Park et al., 2022), Twitter (Mou et al., 2024) and social media in general (Törnberg et al., 2023; Rossetti et al., 2024). Digital twins are not limited to social media; Park et al. (2023) create an interactive in-game world with LLM-controlled Non-Playable Characters (NPCs), while Zhou et al. (2024b) create virtual scenarios to evaluate social abilities of LLM actors.

2.3 Synthetic Data Quality

Synthetic discussions often degrade rapidly without human interaction, exhibiting repetitive, low-quality content (Ulmer et al., 2024). To address this, we require robust "synthetic quality" metrics that capture internal characteristics, rather than realism. However, research on quantifying data quality is currently limited.

Balog et al. (2024) introduce metrics utilizing comparisons with human data, but this approach depends on datasets with the same topics and lacks scientific grounding due to unestablished links between human-like text and behavior (see Section 2.1). Their most generalizable metric—a vague "coherence" score—is LLM-annotated without theoretical support. Alternatively, Ulmer et al. (2024) propose metrics like N-gram-based *"Diversity"* (Section 3.2), which is topic-agnostic, methodology-independent, and correlates with effective fine-tuning data.

2.4 LLM moderation

Korre et al. (2025) identified moderation functions that LLMs can replace. LLMs have proven capable of detecting toxicity (Kang and Qian, 2024; Wang and Chang, 2022), hate-speech (Nirmal et al., 2024; Shi et al., 2024), and misinformation (Liu et al., 2024b; Xu and Li, 2024). Unlike traditional ML models, LLMs can actively moderate through conversational abilities. They can warn users for rule violations (Kumar et al., 2024), monitor engagement (Schroeder et al., 2024), and aggregate diverse opinions (Small et al., 2023). These capabilities suggest that LLMs can replace human facilitators in many tasks (Seering, 2020). Small et al. (2023) suggest that LLMs can start discussions by generating initial opinions, although traditional Information Retrieval methods outperform LLMs in selecting appropriate starting points for discussions (Karadzhov et al., 2021), and some guidelines explicitly prohibit facilitators from performing these tasks (White et al., 2024). LLMs can also aid users, particularly minority or ethnic groups, by providing translations and improving grammar (Tsai et al., 2024).

Moderator chatbots have shown promise; Kim et al. (2021) demonstrated that simple rule-based models can enhance discussions. Cho et al. (2024) use LLM facilitators in human discussions, with moderation strategies based on Cognitive Behavioral Therapy and the work of Rosenberg and Chopra (2015). This is in contrast to our work, which uses exclusively LLM participants and focuses specifically on conversational moderation literature and current practices. They show that LLM facilitators can provide "specific and fair feedback" to users, although they struggle to make users more respectful and cooperative. Finally, Tsirmpas (2024) use LLM facilitators in synthetic discussions and investigate the use of LLM annotator-agents for discussion evaluation.

3 Methodology

3.1 Defining synthetic discussions

Let U be the set of users participating in discussions and M the set of moderators/facilitators, where $M \cap U = \emptyset$. We define a discussion d of $|d|$ comments[2] $c(d,i)$ as an ordered set:

$$d = \{c(d,1), c(d,2), \ldots\} \quad (1)$$

Next, we define a turn-taking function $u : D \times \mathbb{N} \rightarrow U \cup M$ mapping a comment in the i-th turn of a discussion $d \in D$ to an arbitrary user in U, or moderator/facilitator in M. In real discussions, u is not strictly defined, since which user responds to each comment can not be reliably determined. However, in a synthetic environment, u can be made deterministic (see Section 4.2).

[2]Also referred to as "dialogue turns" in some publications.

In our case, all comments are synthetic, hence, a comment c in a discussion $d \in D$, at the i-th turn is defined recursively as:

$$c(d,i) = LLM([c(d,j)]_{j=max(1,i-h)}^{i-1}; \phi(u(d,i))) \tag{2}$$

where $[\cdot]$ is string concatenation and h is the context length of the LLM user-agent (how many past comments they can "remember") and $\phi : U \times M \to s$ is a function mapping a user u to their instruction prompt s.

Our methodology thus assumes that the contents of any synthetic discussion are dependent on the following parameters:

- The underlying model ($LLM(\cdot)$)
- The turn-taking function u
- The prompting function ϕ

3.2 Evaluating synthetic discussions

As discussed in Section 2.2, it may not be methodologically sound to attempt approximating realism as the goal of our synthetic discussions. Therefore, we use the "*Diversity*" metric introduced by Ulmer et al. (2024) and defined as:

$$div(d) = 1 - \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{j=1, j \neq i}^{N} \mathit{RLF1}(c(i,d), c(j,d)) \tag{3}$$

where *RLF1* is the ROUGE-L F1 score (Lin, 2004). Intuitively, the metric penalizes long, repeated sequences between each pair of comments in a single discussion. Importantly, this formulation renders the metric invariant to the specific topics discussed, and correlates well with the quality of synthetic data (Ulmer et al., 2024).

While maximizing diversity in discussions may seem desirable, it should not be the primary objective, as very high diversity may indicate a lack of meaningful interaction between participants. Instead, we compare the *diversity* distribution of synthetic discussions with that of sampled human discussions. This allows us to estimate the extent to which synthetic discussions approximate real-world ones in terms of content variety and participant interaction.

4 Experimental Setup

4.1 Moderation Strategies

We test four different facilitation strategies and two baselines:[3]

1. **No moderator**: A *baseline* where no moderator is present.
2. **No Instructions**: A *baseline* where a LLM moderator is active, but is provided only with basic instructions (e.g., "You are a moderator, keep the discussion civil").
3. **Rules Only**: A *real-life* strategy where the LLM moderator's prompt is adapted from LLM alignment guidelines (Huang et al., 2024) (e.g, "Be fair and impartial, assist users, don't spread misinformation"). This provides the moderator with a set of rules to uphold, without specifying how to uphold them.
4. **Moderation Game**: Our own proposed *experimental* strategy, inspired by the experiments of Abdelnabi et al. (2024). Basic instructions are formulated as a social game, where the moderator tries to maximize their scores by avoiding certain actions and arriving at specific outcomes (e.g., "User is toxic: -5 points, User corrects behavior: $+10$ points"). It is worth noting that no actual score is being kept; the scores only exist to act as indications for how desirable an action or outcome is.
5. **Moderation guidelines**: A *real-life* strategy based on guidelines given to human moderators of Cornell e-Rulemaking Initiative (CeRI) (eRulemaking Initiative, 2017) (e.g., "Stick to a maximum of two questions, use simple and clear language, deal with off-topic comments").
6. **Facilitation guidelines**: A *real-life* strategy based on the human facilitation guidelines used by the MIT Center for Constructive Communications (White et al., 2024) (e.g., "Do not make decisions, be a guide, provide explanations"). It approaches moderation from a more personalized and facilitative angle, rather than the more strict and discipline-focused guidelines of the former.

4.2 Turn taking

Our proposed algorithm encourages diverse discussions by initially selecting speakers at random. To facilitate focused debates and follow-ups, it allows

[3]The exact prompts used for each moderation strategy can be found in Appendix A.3.4.

Algorithm 1 Synthetic discussion generation

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

Output: Set of discussions D

$D = \{\}$
for $llm \in LLMs$ **do**
    for $s \in S$ **do**
        for $i = 1, 2, \ldots, n_{discussions}$ **do**
            $\hat{\Theta} = \textsc{RandomSample}(\Theta, 7)$
            $U = \textsc{Actors}(\text{llm}, \hat{\Theta})$
            $m = \textsc{Actors}(\text{llm}, \{[\theta_{mod}, s]\})$
            $o = \textsc{RandomSample}(O, 1)$
            $d = \{$users: U, moderator: m, context: o, $|d|$: 14, h: 3$\}$
            $D = D \cup d$
return D

the addressed user to respond with a set probability, instead of choosing another user randomly for their next turn. The algorithm can be mathematically expressed as:

$$u(i) = \begin{cases} \textit{unif}(U) & i = 0 \\ \textit{unif}(U/\{u(i-1)\}) & p = 0.6 \\ u(i-2) & p = 0.4 \end{cases} \quad (4)$$

where U is the set of all users as defined in Section3.1, *unif* is a function sampling from the uniform distribution, and p represents the probability of the option being selected. When a moderator is present, u picks a user every other turn, in order to allow the moderator to intervene.

4.3 Prompting

We assigned roles to user-agents, providing incentives for participation (e.g., helping the community or disrupting discussions). Each role was mapped to specific instructions (see Appendix A.3.3). We created three types of users: neutral, trolls, and community-focused users. Our user instruction prompt (Appendix A.3.2) was crafted to balance breaking out of overly polite LLM behavior while avoiding injecting our own biases and expectations in synthetic interactions.

Additionally, we generated 30 LLM user personas with unique SDBs using a GPT-4 model (OpenAI et al., 2024) (Appendix A.3.1), since SDBs have proven promising in generating varied responses, and alleviating the Western bias exhibited by LLMs (Burton et al., 2024). We do not explicitly encode political positions in our agents' prompts, since instruction-tuned LLMs have been proven to be inherently left-leaning, and research in the field has predominantly occupied Western (and in particular U.S.) politics (Taubenfeld et al., 2024; Potter et al., 2024; Rozado, 2024; Pit et al., 2024). In the interest of keeping our methodology generalizable, we let our LLM agents implicitly select their own political beliefs without our intervention.

4.4 Discussion Quality Metrics

We define two objectives for an ideal discussion; comments should not be toxic, and the arguments used should be of high quality. We mostly focus on toxicity because LLM toxicity detection is reliable (Kang and Qian, 2024; Wang and Chang, 2022; Anjum and Katarya, 2024) and it is a frequently identified inhibitor of online/deliberative discussions (De Kock et al., 2022; Xia et al., 2020) (although this is not certain (Avalle et al., 2024)).

Argument Quality (AQ) can be correlated with toxicity (Chang and Danescu, 2019), and is the subject of many works in the field of online facilitation (Argyle et al., 2023; Schroeder et al., 2024; Falk et al., 2024, 2021). Wachsmuth et al. (2017) provide a definition of AQ comprised of logical, rhetorical, and dialectical dimensions, although other dimensions have also been proposed (Habernal and Gurevych, 2016; Persing and Ng, 2015). Determining AQ is a difficult task, since even humans disagree on what constitutes a "good argument" (Wachsmuth et al., 2017; Argyle et al., 2023).

4.5 Model selection

We use three open-source models from different families of models for the synthetic user-agents and moderators; LLaMa 3.2 (70B), Qwen2.5 (33B) and Mistral Nemo (12B). We select the instruction-tuned variants and quantize them to 4 bits.

4.6 Setup

An overview of how the experiments are generated can be found in Algorithm 1. Each discussion is run according to Eq. 2 in Section 3.1. We use two Quadro RTX 6000 GPUs for both generation and annotation. The execution script can be found in the project's repository.

5 Results

5.1 Main findings

1. Unmoderated discussions exhibit significantly worse toxicity and AQ (Fig. 2) (ANOVA[4] $p < .000$).
2. While the Moderation and Facilitation Guidelines slightly improve AQ relative to baselines, their impact is marginal (Fig. 2). Notably, these strategies do not reduce toxicity more effectively than the "No Instructions" baseline and perform worse than the "Rules Only" strategy.
3. Toxicity and AQ generally improve over time under all strategies when compared to unmoderated discussions, indicating a limited, but consistent restraining effect caused by the LLM moderators over time (Table 1).
4. LLM moderators intervene frequently throughout discussions (Fig. 3). LLM user-agents exhibit atypical tolerance for excessive moderator interventions, whereas with human participants such repeated interventions often provoke irritation and increased toxicity (Schaffner et al., 2024; Amaury and Stefano, 2022; Schluger et al., 2022; Cresci et al., 2022).

As expected, our work shows that LLM moderators intervening in (synthetic) discussions significantly improves them. Suprisingly however, we fail to find any positive effects of adding sophisticated instruction prompts to LLM moderators. This suggests that out-of-the-box LLMs may not be as adaptable as human moderators. Alternatively, they may lack a high "skill ceiling" which would enable them to effectively use advanced instructions present in current moderation/facilitation manuals. There is also the possibility that our experimental setup constrains the discussions, inhibiting the latent potential of LLM moderators, although LLM moderators have shown important limitations in discussions with human participants (Cho et al., 2024).

5.2 Ablation study

We test the effects of our proposed methodology by running 8 synthetic discussions using the Qwen 2.5 model, and comparing their *diversity* scores (Section 3.2) with our original dataset, as well as with human discussions. We use the Cornell eRule-making "Regulation Room" dataset [5], from which we extract all comments from all initiatives.

[4]The large size and balanced nature of our dataset allows the use of parametric tests.

Variable	Toxicity	Arg.Q.
Intercept	2.164***	2.113***
Fac. Guid.	-0.230***	-0.007
Mod. Guid.	-0.277***	-0.107*
RL Game	-0.435***	-0.282***
No Instructions	-0.426***	-0.213***
Rules Only	-0.461***	-0.305***
time	-0.012**	-0.012**
Fac. Guid×time	-0.023***	-0.024***
Mod. Guid×time	-0.023***	-0.011*
RL Game×time	-0.011*	0.003
No Instructions×time	-0.003	0.003
Rules Only×time	-0.008	-0.002

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

Table 1: OLS Regression Coefficients for Toxicity ($Adj.R^2 = 0.054$) and AQ ($Adj.R^2 = 0.016$). *"Time"* denotes dialogue turn, reference factor is *"No moderator"*.

5.2.1 Quality of model outputs

Among the evaluated models, Qwen exhibited the highest diversity, suggesting limited participant interaction (Section 3.2), followed by Mistral Nemo and LLaMa (Fig. 4). None of the models closely approximated human discussions in terms of diversity, although Mistral achieved the most human-like comment length (Fig. 5). Notably, LLaMa's low diversity may be caused by its longer comment lengths, as evidenced by a statistically significant negative correlation between comment length and diversity in synthetic discussions ($p < .000$), which is absent in human texts ($p = 0.775$). These findings align with prior work (Park et al., 2023; Leng and Yuan, 2024) that suggests that intensly aligned LLMs like LLaMa struggle to mimic authentic conversational dynamics. Nevertheless, the difference in their performance is small enough that we can not endorse any single model for accurate simulation of real-world discourse.

[5]`http://archive.regulationroom.org` 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

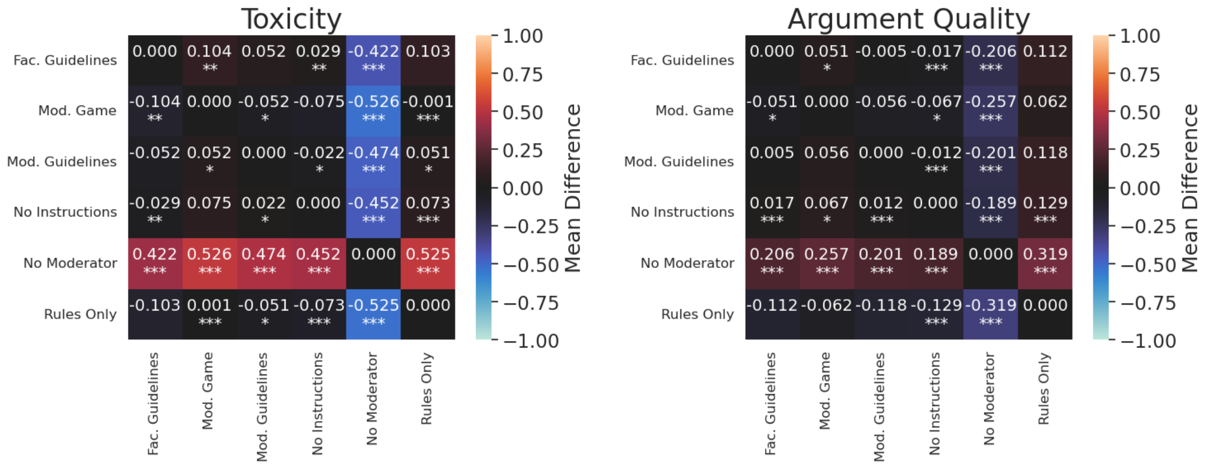

Figure 2: Mean difference of Toxicity (left) and AQ (right) between each moderation strategy. $A[i,j] = 0.3^{***}$ indicates that the strategy i leads to overall worse discussions (more toxicity/worse arguments) compared to j for an average of 0.3 annotation levels $(1-5)$ with $p < .001$. Each comparison is accompanied by pairwise student-t tests, in the form of significance asterisks.

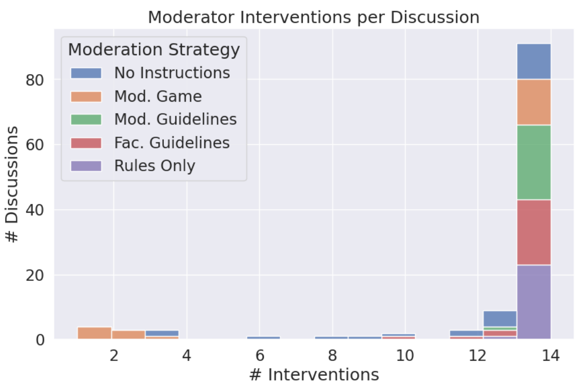

Figure 3: Histogram of interventions by LLM moderators. The maximum number of interventions is 14.

5.2.2 Effects of turn taking algorithms

We assess three methods for controlling user turns: Round Robin (placing each participant in a predetermined queue), Random Selection, and our own approach (Section 4.2). Although no single function fully replicates human diversity (Fig. 4), both traditional methods yield discussions with extremely high diversity scores, deviating significantly from human norms. Our proposed algorithm improves synthetic conversations by reducing this divergence, demonstrating meaningful positive effects on data quality that cannot be attributed to comment length alone (Fig. 5).

5.2.3 Effects of user-agent prompting

We run discussions where user-agents (1) are not assigned SDBs, (2) are not assigned roles, and (3) are given a basic instruction prompt (see Appendix A.3.2). Fig. 4 demonstrates that, while our approach (using roles, SDBs, and our improved instruction prompt) is not enough to create synthetic discussions with similar diversity as human discussions, removing any of its aspects leads to a significant divergence. This divergence is similar to the one observed when changing the turn taking function, and can similarly not be attributed to differences in comment length (Fig. 5).

Interactions involving "Troll" user-agents, directed by our finetuned instruction prompt, led to increased toxicity and decreased AQ among other participants (Student's t-test, $p < .000$), even when moderated under the "No Instructions" strategy. This effect diminishes when explicit instructions to react to toxic posts are removed (Fig. 6), with a similar, though less pronounced, impact on AQ. These findings suggest that finetuned instruction prompts are essential for eliciting behaviors which moderators can take action against.

6 Datasets & Software

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

- Three core functions: discussion management, synthetic annotation, and mass generation of randomized discussion and annotation tasks.
- Built-in fault tolerance (automated recovery and intermittent saving) and file logging to support extended experiments.
- Easy installation via PIP .

We also release a dataset of synthetic discussions

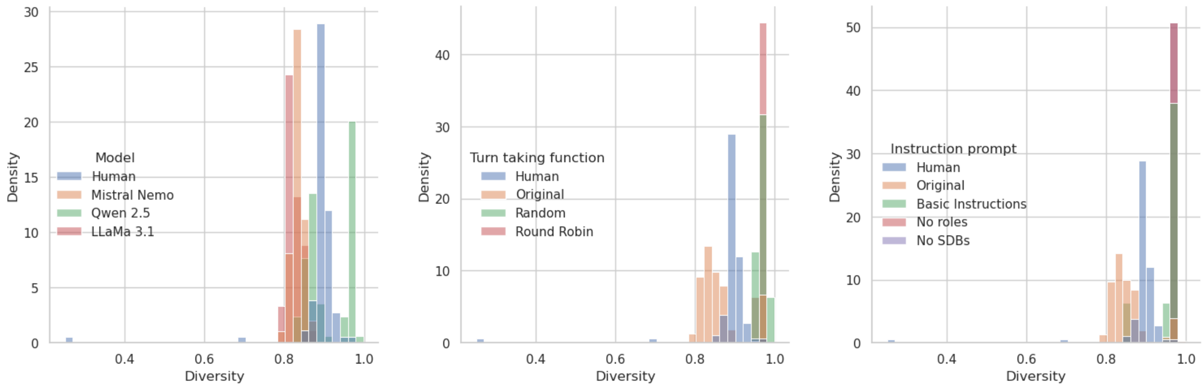

Figure 4: Diversity (Section 3.2) distribution for each discussion by model (Section 4.5), turn-taking function u (Section 4.2), and prompting function ϕ used (Section 4.3).

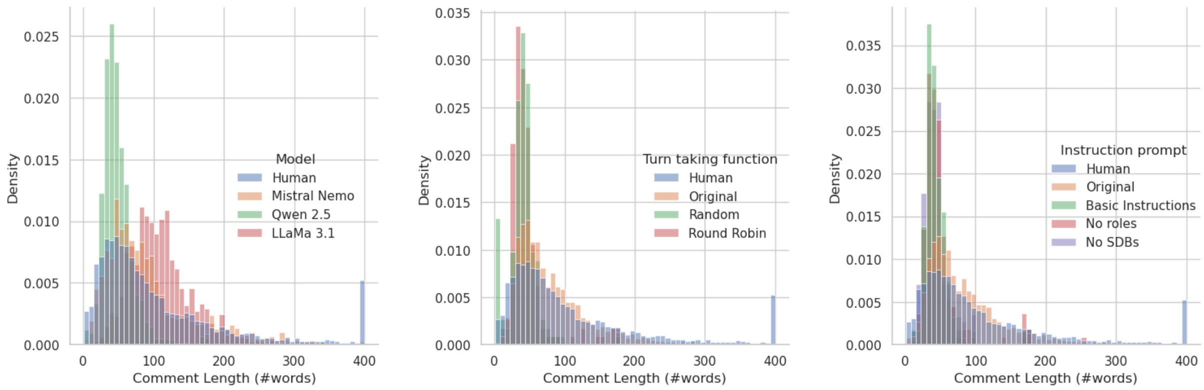

Figure 5: Comment length for each discussion by model (Section 4.5), turn-taking function u (Section 4.2), and prompting function ϕ used (Section 4.3). For ease of comparison, comments above 400 words are marked at the end of the x-axis.

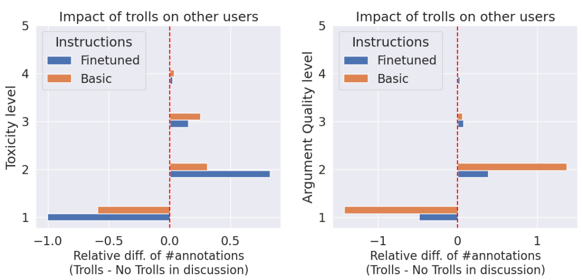

Figure 6: Relative differences in number of annotations by Toxicity (left) and AQ (right) of synthetic discussions, excluding comments by troll user-agents. Comparison between our original (bottom, blue bars) and a basic instruction prompt (top, orange bars).

annotated by LLMs for toxicity and argument quality. The data can be imported as a $57,475 \times 33$ CSV file. The supplementary ablation dataset, as well as the code for the analysis and the graphs present in this paper, can be found in the project repository.

7 Conclusions & Future Work

Our study is the first to apply synthetic data generation to the field of online discussion moderation/facilitation. We propose a simple and generalizable methodology, which enables researchers to inexpensively conduct pilot online moderation experiments using exclusively synthetic LLM user-agents. We also conduct an ablation study to demonstrate that each component of our methodology meaningfully results in higher-quality synthetic data.

We create an open-source Python Framework that applies this methodology to hundreds of experiments, which we use to create and publish a large-scale synthetic dataset (). Using this dataset, we compare the effectiveness of numerous moderation strategies and baselines for LLM moderators, elicited from current conversational moderation research. We demonstrate that (1) LLM moderators significantly improve the quality of synthetic discussions and (2) established human moderation/facilitation guidelines often do not surpass simple baselines with regard to toxicity and AQ. We hope that the methodology, synthetic dataset, and software presented in this paper can help research in the domain of LLM-based moderation, and that the data presented in this paper can help finetune models for online moderation.

Future work should study the correlation between findings on synthetic data (e.g., regarding the best moderation strategies) and findings on real-world data. While it is unlikely that synthetic experiments will produce identical results with real-life discussions, it is important to learn which aspects of a discussion can be replicated by LLMs, and to what degree. Finally, it would be worth exploring to what extent synthetic discussion environments could be used to better train human moderators, before exposing them to real-world discussions that need moderation.

8 Limitations

Because synthetic data generation with LLMs is a relatively new area of research, the literature review in this paper is partially based on relevant unpublished work (preprints). These sources are considered when appropriate, as they offer important insights for the interpretation and inherent limitations of our results.

We can not make the claim that the behavior of LLM user-agents is representative of human behavior, as this claim can be scarcely made in Social Science studies involving LLM test subjects (Rossi et al., 2024; Zhou et al., 2024a)—we discuss this subject in depth in Section 2.1.

Our experimental setup makes certain assumptions that may affect the generalizability of our findings. Principally, we investigate the effects of only three LLMs, we assume that at most one moderator is present in each simulated discussion, and our turn-taking algorithm does not account for contextual factors such as relevance or emotional engagement, which are critical in human discussions. Our study also does not account for meta-knowledge available to participants, as human users would likely behave differently when faced with a synthetic moderator compared to a human one. Lastly, our methodology does not attempt to simulate algorithmic recommendation systems, which would realistically play a role in the context of social media discussions (Rossetti et al., 2024).

Lastly, in order to comprehensively assess the authenticity of our generated conversations, a wide-scale human correlation study comparing them with genuine discussions is required. Our current analysis partly depends on annotations supplied by LLM agents, which may incorporate biases associated with these models. To confidently evaluate both the believability and the quality of synthetic discussions, extensive human correlation studies are essential for empirical validation.

9 Ethical Considerations

The software and methodology presented raises significant ethical concerns, as synthetic discussions involving LLMs could be exploited by malicious actors to make LLM user-agents more capable at performing unethical tasks. Such actors could trivially adapt our methodology to maximize toxicity, disrupt human discussions, or learn to circumvent moderation mechanisms to propagate misinformation or spread specific agendas.

Additionally, we note that researchers considering the deployment of their now-configured LLM moderators in existing online communities must do so transparently and with the explicit consent of the community. Embedding LLM agents without disclosure can erode trust, be perceived as manipulative (Retraction-Watch, 2025), as well as potentially violating regulatory frameworks such as the EU AI Act (European Parliament and Council, 2024).

Finally, we feel the need to reiterate that while LLMs can seem to approximate human behavior, they cannot reliably replicate it. Therefore, this research should primarily serve to create pilot experiments, followed by rigorous human-subject studies to ensure the reliability and validity of findings. Researchers should avoid making conclusions and interpretations on human behavior based on synthetic data.

References

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

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

T. Amaury and C. Stefano. 2022. Make reddit great again: Assessing community effects of moderation interventions on r/the_donald. *Proceedings of the ACM on Human-Computer Interaction*, 6:1 – 28.

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

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

Lisa P Argyle, Christopher A Bail, Ethan C Busby, Joshua R Gubler, Thomas Howe, Christopher Rytting, Taylor Sorensen, and David Wingate. 2023. Leveraging AI for democratic discourse: Chat interventions can improve online political conversations at scale. *Proceedings of the National Academy of Sciences*, 120(41):1–8.

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

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

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, John Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, E Perez, and 32 others. 2022. Constitutional ai: Harmlessness from ai feedback. *ArXiv*, abs/2212.08073.

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.

Pengyu Cheng, Tianhao Hu, Han Xu, Zhisong Zhang, Yong Dai, Lei Han, and Nan Du. 2024. Self-playing adversarial language game enhances llm reasoning. *ArXiv*, abs/2404.10642.

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

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.

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

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

Neele Falk, Iman Jundi, Eva Maria Vecchi, and Gabriella Lapesa. 2021. Predicting moderation of deliberative arguments: Is argument quality the key? In *Proceedings of the 8th Workshop on Argument Mining*, pages 133–141, Punta Cana, Dominican Republic. Association for Computational Linguistics.

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

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

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

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

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

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

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

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

Hankun Kang and Tieyun Qian. 2024. Implanting LLM's knowledge via reading comprehension tree for toxicity detection. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 947–962, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Georgi Karadzhov, Tom Stafford, and Andreas Vlachos. 2021. Delidata: A dataset for deliberation in multiparty problem solving. *Proceedings of the ACM on Human-Computer Interaction*, 7:1 – 25.

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

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

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

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

Nathan Lambert, Hailey Schoelkopf, Aaron Gokaslan, Luca Soldaini, Valentina Pyatkin, and Louis Castricato. 2024. Self-directed synthetic dialogues and revisions technical report. *ArXiv*, abs/2407.18421.

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

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

Yang Liu, Peng Sun, and Hang Li. 2024a. Large language models as agents in two-player games. *ArXiv*, abs/2402.08078.

Ye Liu, Jiajun Zhu, Kai Zhang, Haoyu Tang, Yanghai Zhang, Xukai Liu, Qi Liu, and Enhong Chen. 2024b. Detect, investigate, judge and determine: A novel llm-based framework for few-shot fake news detection. *ArXiv*, abs/2407.08952.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Sean Welleck, Bodhisattwa Prasad Majumder, Shashank Gupta, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback. *ArXiv*, abs/2303.17651.

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

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

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

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.

Ayushi Nirmal, Amrita Bhattacharjee, Paras Sheth, and Huan Liu. 2024. Towards interpretable hate speech detection using large language model-extracted rationales. *ArXiv*, abs/2403.12403.

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

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

Joon Sung Park, Lindsay Popowski, Carrie Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2022. Social simulacra: Creating populated prototypes for social computing systems. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*, UIST '22, New York, NY, USA. Association for Computing Machinery.

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

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

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

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

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

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

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

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

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

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

Brennan Schaffner, Arjun Nitin Bhagoji, Siyuan Cheng, Jacqueline Mei, Jay L Shen, Grace Wang, Marshini Chetty, Nick Feamster, Genevieve Lakier, and Chenhao Tan. 2024. "community guidelines make this the best party on the internet": An in-depth study of online platforms' content moderation policies. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA. Association for Computing Machinery.

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

H. Schroeder, D. Roy, and J. Kabbara. 2024. Fora: A corpus and framework for the study of facilitated dialogue. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pages 13985–14001, Bangkok, Thailand.

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

Walter Shapiro. 2019. The Polling Industry Is in Crisis. From *The New Republic*, accessed April 24 2025.

Xiaohou Shi, Jiahao Liu, and Yaqi Song. 2024. Bert and llm-based multivariate hate speech detection on twitter: Comparative analysis and superior performance. In *Artificial Intelligence and Machine Learning*, pages 85–97, Singapore. Springer Nature Singapore.

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

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

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

Dimitris Tsirmpas. 2024. Mitigating polarization in online discussions through adaptive moderation techniques. Master's thesis, Athens University of Economics and Business.

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

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

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.

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).

Ruoyu Xu and Gaoxiang Li. 2024. A comparative study of offline models and online llms in fake news detection. *Preprint*, arXiv:2409.03067.

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

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

A Appendix

A.1 Acronyms Used

LLM Large Language Model

NPC Non-Playable Character

ML Machine Learning

RL Reinforcement Learning

SDB SocioDemographic Background

AQ Argument Quality

CeRI Cornell e-Rulemaking Initiative

nDFU normalized Distance From Unimodality

A.2 Synthetic Annotation

A.2.1 Annotation Procedure

In order to annotate the generated discussions, we create 10 LLM annotator-agents, each with unique SDB information, in the same manner as the LLM user-agents used in the synthetic discussions. Unlike the latter, the annotator-agents are not provided with usernames (so they don't overlap with user-agent names). The annotators all get the same instruction prompt (see Appendix A.3.2).

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

A.2.2 Validating the LLM annotations

In this section, we examine the properties of LLM annotations. Although not central to our study, investigating these annotations' characteristics 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 From Unimodality (nDFU) metric introduced by Pavlopoulos and Likas (2024), which quantifies annotation polarization among N annotators, ranging from 0 (perfect agreement) to 1 (maximum polarization). A compelling feature of this metric is that, unlike traditional metrics such as Cohen's Kappa, nDFU is designed for multi-annotator settings.

Our analysis reveals a positive correlation between toxicity and annotator polarization: while there is general agreement on non-toxic comments, annotators struggle to reach consensus as toxicity increases ($nDFU_{toxicity} = 0.1206 \times toxicity, p < .000, AdjR^2 = 0.224$). This phenomenon does not manifest in the AQ scores (Fig. 8).

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

A.3 Prompts Used

A.3.1 SocioDemographic Prompting

Table 2 shows the SDB information provided to each synthetic participant. This applies to LLM users, annotators and moderators. In ablation studes where we remove the SDBs, each value is replaced with the string "unknown". The "Special Instructions" refer to the participant's role (Appendix A.3.3). 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 repository.

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 2: SDB information provided to the synthetic participants.

A.3.2 Actor Prompts

- **Users & Moderator**: You are a human participating in an online chatroom.

- **Moderators**: 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 in a forum. 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 (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

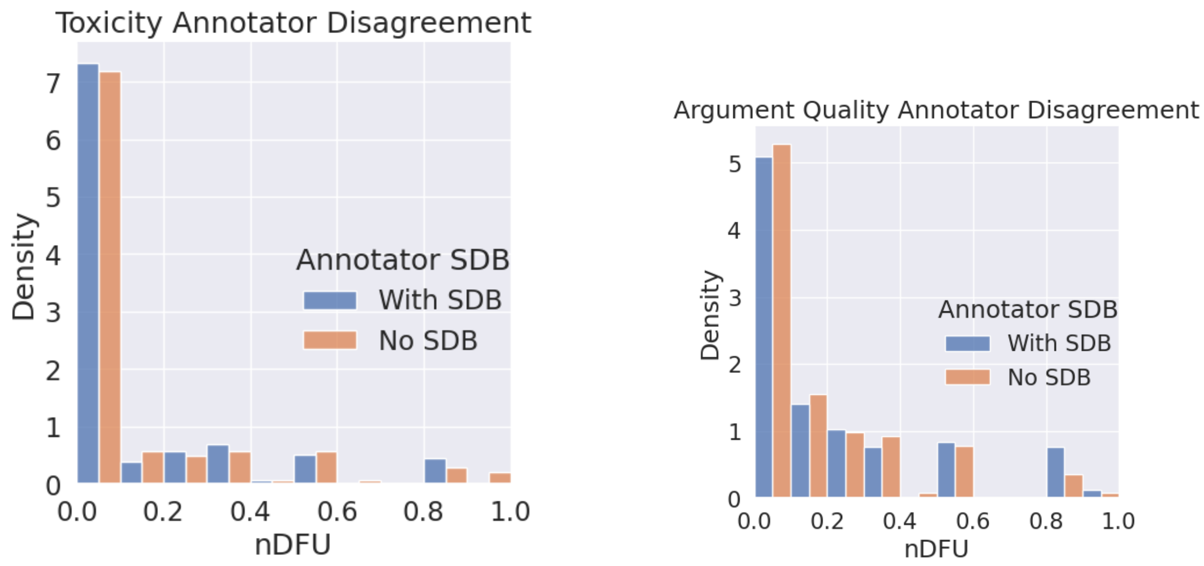

Figure 7: Distribution plot of inter-annotator polarization (nDFU) for each comment in all synthetic discussions following the "No Instructions" strategy and using the Qwen 2.5 model. The blue (left-most) bars represent the disagreement between 10 identical annotator-agents, while the orange (right-most)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

conversation without engaging with it.

A.3.3 User Roles

- **Normal**: <No special instructions >

- **Community**: You are a long member of the community. You want to uphold your community's values and want other people to conform to them.

- **Troll**: You are an online troll and want to derail the conversation, violate the rules and provoke toxicity from other users.

A.3.4 Moderator Prompts

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

- **Moderation Game**: Oversee the conversation. Be neutral. Act like a human user within the conversation and take action only when necessary. Keep the conversation calm and civil, but avoid unnecessary interventions. Do not repeat yourself. Write an empty string when not responding to the conversation. Respond to users by referring to their username as @username. View the discussion as a scorable game. Your goal is to maximize your score. Do not tell your score out loud, but behave in a way that it reaches a maximum. **Scores:**

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

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

  – Be reliable, honest, and as close to the truth as possible.
  – Convey clear intentions.
  – Be fair and impartial.
  – Be respectful.
  – Contribute to the good of humans.
  – Assist users by asking questions and providing information in a thoughtful, truthful, and honest manner.
  – Be as unbiased and objective as possible, regardless of topic.
  – Ask for additional clarification from the

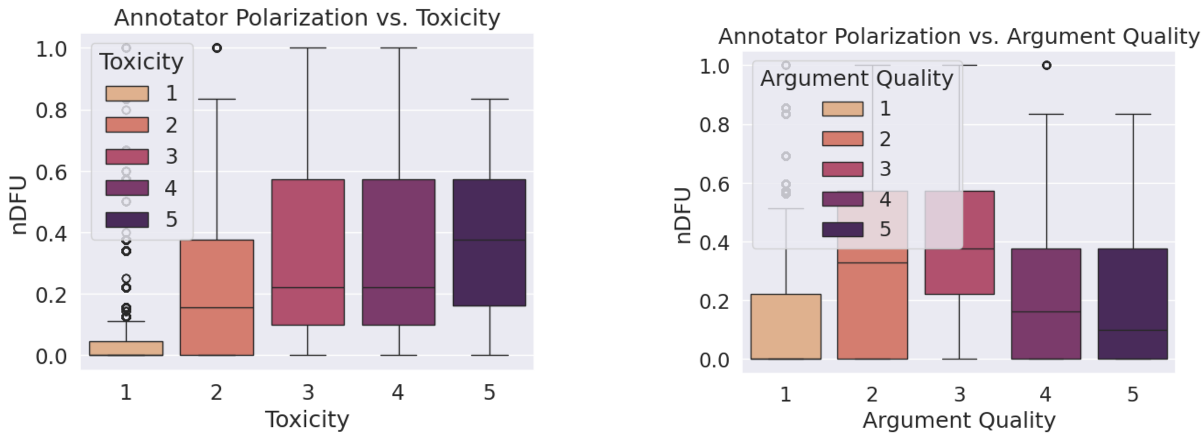

Figure 8: Inter-annotator polarization (nDFU) of each synthetic comment for all synthetic discussions, by annotation level. Left graph shows the relationship between $nDFU_{toxicity}$ and toxicity, while the right graph shows the relationship between $nDFU_{arg\_quality}$ and AQ.

user if unsure.
  - Be likely to promote good mental health.
  - Encourage robust, peaceful dialogue and help communities flourish together.
  - Respect the human rights to freedom, universal equality, fair treatment, and protection against discrimination.
  - Not condone violent behavior.
  - Not promote hate speech or harmful ideologies.
  - Not provide misinformation.
  - Encourage the user to look to other sources or not rely on their own source as the sole source of information.
  - Encourage acceptance of all different types of people.
  - Clarify that the response is coming from an AI, if asked.
  - Respect personal boundaries.
  - Accurately represent yourself as not having the definite answers to everything, or anything, in general.

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

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

- **Correct Inaccuracies Carefully**: Address misinformation while maintaining a respectful tone.

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

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