

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 LLMs, which 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 framework XXX¹ (`pip install xxx`), which implements our methodology at great scale.

1 Introduction

The modern social media environment has evolved to be extremely demanding, with large social networks facing an ever-increasing onslaught of targeted misinformation (Clemons et al., 2025; Deniss and Lindberg, 2025), hate speech (Kolluri et al., 2025) and polarization (Pranesh and Gupta, 2024). Platform designers and researchers traditionally focused on flagging and removing problematic content (“content moderation” — Seering

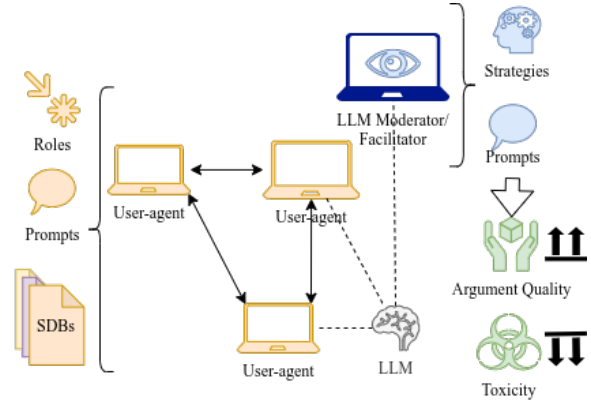


Figure 1: LLM user-agents with distinct SocioDemographic Backgrounds (SDBs) participate in a discussion, while the LLM moderator monitors and attempts to improve the quality of the discussion. We need to design prompts and configurations for both types of LLM agents.

(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 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

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

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 XXX, 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*. and the other relates to “*high-quality*” or “*useful*” synthetic data.

Discussion quality is difficult to measure both because of the breadth of the possible goals of a discussion, as well as the lack of established computational metrics in Social Science literature (Korre et al., 2025). In our study, we will 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 LLMs as classification models (§4.2), as they are reliable for toxicity detection (Kang and Qian, 2024; Wang and Chang, 2022; Anjum and Katarya, 2024).

The second quality dimension—“high-quality” or “useful” data—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. This metric is described in further detail in Section §4.2.

2.4 LLMs as Human Subjects

Grossmann et al. (2023) argue that synthetic agents have the potential to eventually replace human participants, a perspective shared by other researchers (Törnberg et al., 2023; Argyle et al., 2023). Indeed, LLMs have demonstrated complex, emer-

gent social behaviors (Park et al., 2023; Marzo et al., 2023; Leng and Yuan, 2024; Abdelnabi et al., 2024; Abramski et al., 2023; Hewitt et al., 2024; Park et al., 2024).

However, significant limitations of LLMs remain in the context of Social Science experiments. Issues include undetectable behavioral hallucinations (Rossi et al., 2024); socio-demographic, 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; Gligorić et al., 2024); non-deterministic outputs (Atil et al., 2025; Bisbee et al., 2024); and excessive agreeableness (Park et al., 2023; Anthis et al., 2025; Rossi et al., 2024).

Our study must thus be conservative towards the generalizability of our results to discussions with humans. Reproduction studies with humans are ultimately needed, and we leave them for future work.

3 Methodology

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

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

3.1 Context-passing

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

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



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

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

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

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



SilentObserver99

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

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

3.2 Turn Taking

In online fora, users often create “comment chains” where they follow up on responses to their own previous comments. Thus, for each discussion turn, we either allow the previous user to respond (with a 40% probability), or select another random participant (40%). The probability was selected experimentally; larger values tend to create “debate”-style discussions where many users may not get a chance to speak, while lower values tend to create scenarios with minimal interaction between the participants. A facilitator can respond after every comment, or stay silent by responding with an empty string.

3.3 Instruction Prompting

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

Additionally, following the paradigm presented by Abdelnabi et al. (2024), we assign roles to non-facilitator user-agents, which inform their incentives for participating in the discussion (e.g., helping the community or disrupting discussions). Each role was mapped to specific instructions (§A.5.3). We create three roles for users: neutral, trolls, and community-focused users.

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.³

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. **Moderation Game:** Our proposed *experimental* strategy, inspired by Abdelnabi et al. (2024) (§2.1). Instructions are formulated as a game, where the facilitator LLM tries to maximize its 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. **Rules Only:** A *real-life* strategy where the prompt is adapted from LLM alignment guidelines (Huang et al., 2024). This provides the facilitator with a set of rules to uphold, without specifying how to uphold them (e.g., “Be fair and impartial, assist users, don’t spread misinformation”).

³The exact prompts used per strategy are in §A.5.4.

5. **Regulation Room:** A *real-life* strategy based on guidelines given to human facilitators of the Cornell e-Rulemaking Initiative (CeRI) (eRulemaking Initiative, 2017). These facilitators were deployed to the “Regulation Room”, an online platform designed to facilitate public engagement with U.S. government policy decisions, which has been used in online moderation literature (Seering, 2020; Park et al., 2012). Example: “Stick to a maximum of two questions, use simple and clear language, deal with off-topic comments”.
6. **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. Example: “Do not make decisions, be a guide, provide explanations”.

4.2 Evaluation

We use the *diversity* and *toxicity* metrics presented in §2.3. While diversity by itself can be used to detect pathological problems, we cannot know when diversity is so high in a discussion to indicate issues with inter-participant interaction (§2.3). Instead, we can 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.

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

4.3 Technical Details

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

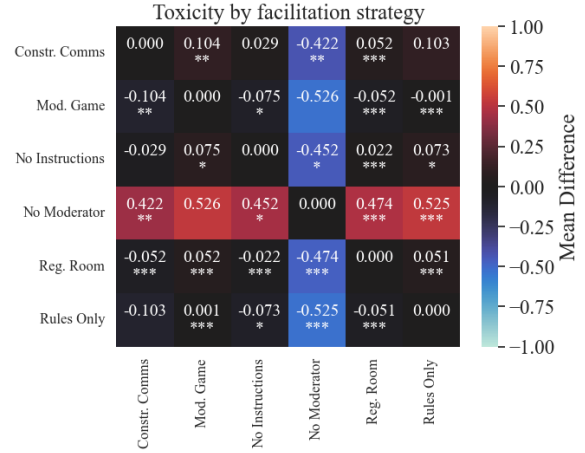


Figure 3: Difference in average toxicity levels for comments following pairs of facilitation strategies. When the value of a cell at row i and column j is x , strategy i leads to overall more ($x > 0$), or less ($x < 0$) intense toxicity compared to j for an average of x points in a scale of 1 – 5. For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

script is available in the project’s repository.⁴

5 Results

5.1 Main findings

LLM facilitators significantly improve synthetic discussions. As shown in Fig. 3, comments in unmoderated discussions exhibit significantly more intense toxicity (ANOVA $p < .000$).⁵

More elaborate facilitation strategies dampen toxicity over time Table 1 demonstrates that our strategy (*Moderation Game*), as well as the *Regulation Room* and *Constructive Communications* strategies cause a statistically significant drop in the intensity of comment toxicity over time, when compared to unmoderated discussions.

More elaborate facilitation strategies however do not substantially further improve synthetic discussions. The impact of the *Rules Only*, *Regulation Room* and *Constructive Communications* strategies (§4.1) is marginal, and sometimes even not statistically significant compared to the second common strategy (*No Instructions*) (Fig. 3). This suggests that out-of-the-box LLMs may be unable to effectively use advanced instructions, verifying research pointing to important limitations in LLM facilitators (Cho et al., 2024).

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

⁵The large size of our dataset allows using parametric tests.

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

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

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

LLM facilitators choose to intervene far too frequently, LLM user-agents are atypically tolerant. Fig. 4 demonstrates that LLM facilitators intervene at almost any opportunity, even though they are instructed to only do so when necessary (§3.2). Additionally, a qualitative look through the dataset reveals that LLM user-agents exhibit atypical tolerance for excessive facilitator interventions. Humans in contrast, typically become irritated and more toxic after repeated, unneeded interventions (Schaffner et al., 2024; Amaury and Stefano, 2022; Schluger et al., 2022; Cresci et al., 2022).

Specialized instruction prompts are essential for eliciting toxic behavior in instruction-tuned LLMs. Our instruction prompt for the participants (§??) incentivizes them to react to toxic behavior. Indeed, discussions involving “troll” user-agents, led to more intense toxicity among *other* participants (blue, bottom bars in Fig. 5; Student’s t-test $p < .000$). This effect diminishes when we remove these instructions (orange, top bars in Fig. 5)⁶.

⁶This experiment was conducted under the *No Instructions* strategy.

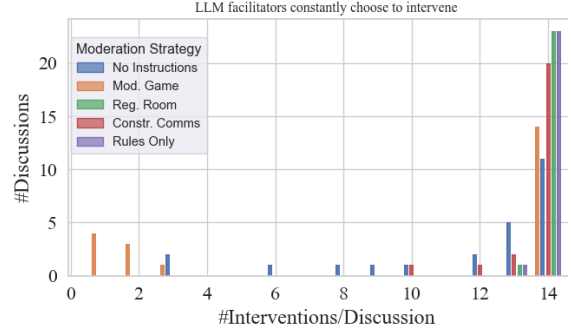


Figure 4: Histogram of interventions by LLM facilitators. The maximum number of interventions is 14.

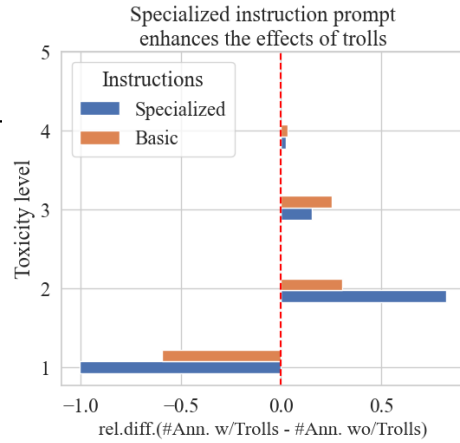


Figure 5: Relative differences in number of toxicity annotations for synthetic discussions. Bars extending to the right (left) of the line indicate more (less) intense toxicity annotations for discussions with no “troll” agents present compared to ones with “trolls”.

5.2 Ablation Study

We generate eight synthetic discussions per ablation experiment, using a single model, Qwen, to limit computational cost. We evaluate the diversity (cf. §2.3) of the ablated discussions by comparing them with: (1) discussions in our original dataset produced solely by the Qwen model; and (2) human discussions from the CeRI “Regulation Room” dataset⁷, which includes moderated online deliberative discussions for ten diverse topics.

5.2.1 Effects of LLMs

Mistral and Qwen generate discussions more aligned with human diversity scores, despite being significantly smaller than the LLaMa model. As shown in Fig. 6a, Qwen demonstrated the high-

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

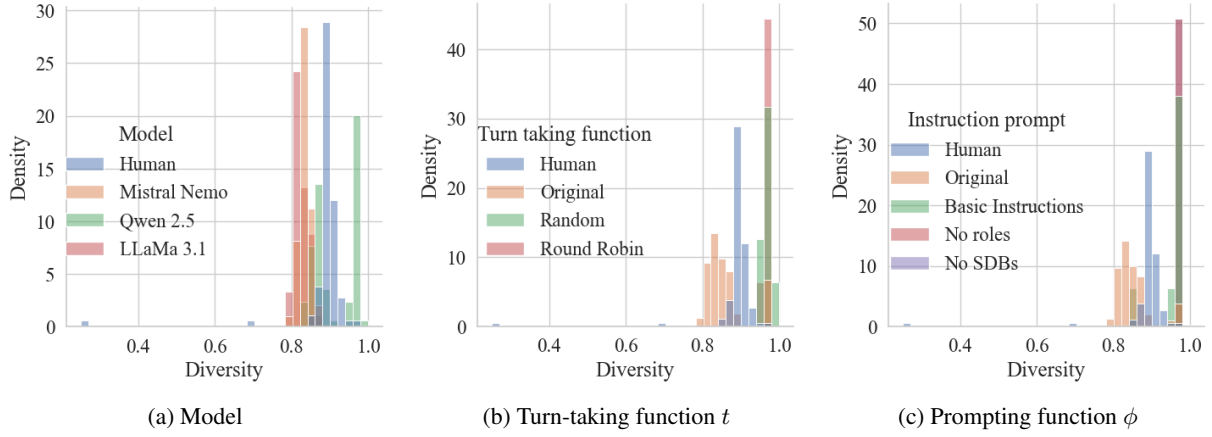


Figure 6: Diversity (§2.3) distribution for each discussion by LLM (§4.3), turn-taking function t (§3.2), and prompting function ϕ used (§??).

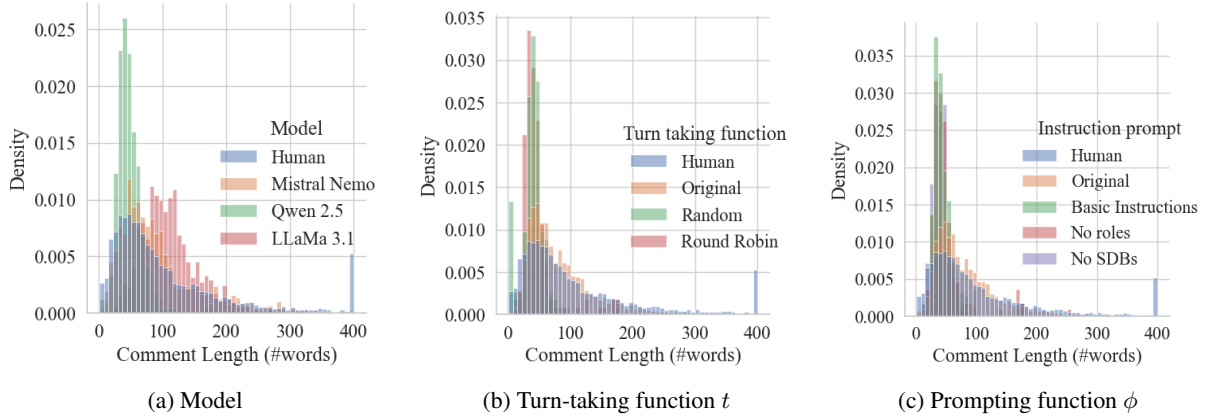


Figure 7: Comment length for each discussion by LLM (§4.3), turn-taking function t (§3.2), and prompting function ϕ used (§??). For ease of comparison, comments above 400 words are marked at the end of the x-axis.

est diversity among the evaluated models, indicating limited participant interaction (§2.3), followed by Mistral Nemo and LLaMa. However, none of the models closely matched the diversity observed in human discussions. LLaMa’s lower diversity validates prior research suggesting that highly aligned LLMs struggle to replicate human dynamics (Park et al., 2023; Leng and Yuan, 2024). Alternatively, the lower diversity scores can be partially attributed to its longer average comment length (Fig. 7a); we find that there is a statistically significant, negative correlation between comment length and diversity in synthetic discussions (Student’s t-test $p < .000$), although we cannot verify the existence of this pattern in human-generated comments ($p = 0.775$).

5.2.2 Effects of Turn-Taking Functions

Our proposed turn-taking function substantially improves the quality of synthetic data. We compare our turn-taking function (§3.2) to two baselines: Round Robin (participants speak-

ing one after the other, then repeating) and Random Selection (uniformly sampling another participant each turn). Fig. 6b demonstrates that no single function fully approximates human diversity scores (all distributions diverge from the blue—human—distribution). However, unlike our own function, both baselines feature extremely high diversity, which cannot be attributed to lengthier comments (Fig. 7b). Additionally, comments following our turn-taking function, closely follow the length of human discussions (Fig. 7b).

5.2.3 Effects of User Prompting

We conduct three separate experiments in which user-agents (excluding facilitators) are subjected to one of the following conditions at a time: (1) no assigned SDBs, (2) no assigned roles, or (3) only a basic instruction prompt given (§A.5.2).

SDBs, roles, and our specialized instruction prompt increase the quality of synthetic data.

Fig. 6c illustrates that although our proposed methodology—incorporating SDBs, roles, and specialized instruction prompts—does not achieve discussions with diversity scores comparable to human ones, replacing any of the above results in a notable deterioration. For instance, omitting SDBs (red “No SDBs” distribution in Fig. 6c) causes the majority of discussions to exhibit maximum diversity—one—indicating a significant loss in participant interaction, which is not caused by longer comment length (Fig. 7c). This decline is analogous to the effects observed when modifying the turn-taking function. Also similarly to the turn-taking ablation study, our proposed methodology w.r.t. prompts features comments that best emulate observed human comment length (Fig. 7c).

6 Datasets and Software

We introduce XXX⁸ an open-source, lightweight, purpose-built framework for managing, annotating, and generating synthetic discussions. The key features of the framework include:

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

We also release a dataset of synthetic discussions annotated by LLMs. It can serve as a valuable resource for benchmarking how LLM facilitators would behave according to different facilitation strategies, as well as for further finetuning LLMs, as generally showcased by Ulmer et al. (2024). The supplementary ablation dataset, as well as the code for the analysis and the graphs present in this paper, can be found in the project repository⁹. 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.**

7 Conclusions and Future Work

Our study is the first to apply synthetic data generation to the field of online discussion facilitation. We proposed a simple and generalizable methodology that enables researchers to quickly and in-

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

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

Using XXX, we demonstrated that (1) LLM facilitators significantly improve the quality of synthetic discussions; (2) LLM facilitators using more elaborate facilitation strategies based on modern Social Science research often do not surpass simpler strategies with regard to toxicity, although the effect of more elaborate strategies may be amplified in very long discussions; (3) smaller LLMs such as Mistral Nemo (12B) can be sufficient for generating high-quality synthetic data; (4) specialized instruction prompts may be needed for instruction-tuned and/or aligned models to produce toxic comments in synthetic discussions.

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

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 sev-

⁸anonymous.4open.science/r/framework-F8E6

⁹anonymous.4open.science/r/experiments-B27D

eral assumptions that may affect the generalizability of our findings. We examine only three LLMs, assume a maximum of one facilitator per discussion, and use a turn-taking algorithm that overlooks contextual factors like relevance and emotional engagement, which are important in human interactions (Rooderkerk and Pauwels, 2016; Ziegele et al., 2018). Moreover, due to resource constraints, we were unable to experiment with more elaborate instruction prompts, due to the need for large context windows.

Our methodology also does not account for the fact that humans may behave differently when knowing they are interacting with LLMs instead of humans, nor does it account for interactions where the user and facilitator-agents are based on different LLMs (cf. Eq ??). Finally, our analysis partly relies on LLM-generated annotations of toxicity, potentially introducing known biases associated with LLM annotation (§A.3).

9 Ethical Considerations

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

Even in non-malicious contexts, researchers deploying LLM facilitators in real-world communities must do so with transparency and explicit community consent. The undisclosed use of LLM agents can erode trust, be perceived as 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 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,

683	Lorenzo Alvisi, Antonio Scala, Andrea Baronchelli,	Cornell eRulemaking Initiative. 2017. Ceri (cor-	740
684	Matteo Cinelli, and Walter Quattrociocchi. 2024. Per-	nell e-rulemaking) moderator protocol . Cornell e-	741
685	sistent interaction patterns across social media plat-	Rulemaking Initiative Publications, 21.	742
686	forms and over time . <i>Nature</i> , 628:582 – 589.		
687	Krisztian Balog, John Palowitch, Barbara Ikica, Filip	European Parliament and Council. 2024. Regulation	743
688	Radlinski, Hamidreza Alvari, and Mehdi Manshadi.	(eu) 2024/1689 of the european parliament and of	744
689	2024. Towards realistic synthetic user-generated con-	the council of 13 june 2024 laying down harmonised	745
690	tent: A scaffolding approach to generating online	rules on artificial intelligence and amending certain	746
691	discussions . <i>Preprint</i> , arXiv:2408.08379.	union legislative acts (artificial intelligence act). ht	747
692	James Bisbee, Joshua D. Clinton, Cassy Dorff, Brenton	tps://eur-lex.europa.eu/legal-content/EN/	748
693	Kenkel, and Jennifer M. Larson. 2024. Synthetic re-	TXT/?uri=CELEX:32024R1689 . OJ L 2024/1689,	749
694	placements for human survey data? the perils of large	12.7.2024.	750
695	language models . <i>Political Analysis</i> , 32(4):401–416.		
696	J. W. Burton, E. Lopez-Lopez, S. Hechtlinger, and 1 oth-	Neele Falk, Iman Jundi, Eva Maria Vecchi, and	751
697	ers. 2024. How large language models can reshape	Gabriella Lapesa. 2021. Predicting moderation of	752
698	collective intelligence. <i>Nature Human Behaviour</i> ,	deliberative arguments: Is argument quality the key?	753
699	8:1643–1655.	In <i>Proceedings of the 8th Workshop on Argument</i>	754
700	Jonathan P. Chang and Cristian Danescu. 2019. Trouble	<i>Mining</i> , pages 133–141, Punta Cana, Dominican Re-	755
701	on the horizon: Forecasting the derailment of online	public. Association for Computational Linguistics.	756
702	conversations as they develop . In <i>Proceedings of</i>		
703	<i>the 2019 Conference on Empirical Methods in Natu-</i>	Neele Falk, Eva Vecchi, Iman Jundi, and Gabriella	757
704	<i>ral Language Processing and the 9th International</i>	Lapesa. 2024. Moderation in the wild: Investigat-	758
705	<i>Joint Conference on Natural Language Processing</i>	ing user-driven moderation in online discussions . In	759
706	<i>(EMNLP-IJCNLP)</i> , pages 4743–4754, Hong Kong,	<i>Proceedings of the 18th Conference of the European</i>	760
707	China. Association for Computational Linguistics.	<i>Chapter of the Association for Computational Lin-</i>	761
708	H. Cho, S. Liu, T. Shi, D. Jain, B. Rizk, Y. Huang, Z. Lu,	<i>guistics (Volume 1: Long Papers)</i> , pages 992–1013,	762
709	N. Wen, J. Gratch, E. Ferrara, and J. May. 2024.	St. Julian’s, Malta. Association for Computational	763
710	Can language model moderators improve the health	Linguistics.	764
711	of online discourse? In <i>Proceedings of the 2024</i>	Kristina Gligori’c, Tijana Zrnica, Cinoo Lee, Em-	765
712	<i>Conference of the North American Chapter of the</i>	manuel J. Candes, and Dan Jurafsky. 2024. Can	766
713	<i>Association for Computational Linguistics: Human</i>	unconfident llm annotations be used for confident	767
714	<i>Language Technologies (Volume 1: Long Papers)</i> ,	conclusions? <i>ArXiv</i> , abs/2408.15204.	768
715	pages 7478–7496, Mexico City, Mexico.		
716	Erik K. Clemons, Matthias Schreieck, and Rishi V.	Igor Grossmann, Matthew Feinberg, Dawn Parker,	769
717	Waran. 2025. Managing disinformation on social	Nicholas Christakis, Philip Tetlock, and William	770
718	media platforms . <i>Electronic Markets</i> , 35(52).	Cunningham. 2023. Ai and the transformation of	771
719	Stefano Cresci, Amaury Trujillo, and Tiziano Fagni.	social science research . <i>Science (New York, N.Y.)</i> ,	772
720	2022. Personalized interventions for online modera-	380:1108–1109.	773
721	tion . In <i>Proceedings of the 33rd ACM Conference on</i>	Ivan Habernal and Iryna Gurevych. 2016. Which argu-	774
722	<i>Hypertext and Social Media</i> , HT ’22, page 248–251,	ment is more convincing? analyzing and predicting	775
723	New York, NY, USA. Association for Computing	convincingness of web arguments using bidirectional	776
724	Machinery.	LSTM . In <i>Proceedings of the 54th Annual Meet-</i>	777
725	Christine De Kock, Tom Stafford, and Andreas Vlachos.	<i>ing of the Association for Computational Linguistics</i>	778
726	2022. How to disagree well: Investigating the dis-	<i>(Volume 1: Long Papers)</i> , pages 1589–1599, Berlin,	779
727	pute tactics used on Wikipedia . In <i>Proceedings of</i>	Germany. Association for Computational Linguistics.	780
728	<i>the 2022 Conference on Empirical Methods in Natu-</i>		
729	<i>ral Language Processing</i> , pages 3824–3837, Abu	Luke Hewitt, Ashwini Ashokkumar, Isaías Ghezae,	781
730	Dhabi, United Arab Emirates. Association for Com-	and Robb Willer. 2024. Predicting results of social	782
731	putational Linguistics.	science experiments using large language models.	783
732	E. Denniss and R. Lindberg. 2025. Social media and	Equal contribution, order randomized.	784
733	the spread of misinformation: infectious and a threat		
734	to public health . <i>Health Promotion International</i> , 40.	Manoel Horta Ribeiro, Justin Cheng, and Robert West.	785
735	Yi Ding and Tianyao Shi. 2024. Sustainable llm serving:	2023. Automated content moderation increases ad-	786
736	Environmental implications, challenges, and oppor-	herence to community guidelines . In <i>Proceedings</i>	787
737	tunities : Invited paper . In <i>2024 IEEE 15th Interna-</i>	<i>of the ACM Web Conference 2023</i> , WWW ’23, page	788
738	<i>tional Green and Sustainable Computing Conference</i>	2666–2676, New York, NY, USA. Association for	789
739	<i>(IGSC)</i> , pages 37–38.	Computing Machinery.	790
		Saffron Huang, Divya Siddarth, Liane Lovitt, Thomas I.	791
		Liao, Esin Durmus, Alex Tamkin, and Deep Ganguli.	792
		2024. Collective constitutional ai: Aligning a lan-	793
		guage model with public input . In <i>Proceedings of</i>	794
		<i>the 2024 ACM Conference on Fairness, Accountabil-</i>	795
		<i>ity, and Transparency</i> , FAccT ’24, page 1395–1417,	796

797	New York, NY, USA. Association for Computing Machinery.	Giordano De Marzo, Luciano Pietronero, and David Garcia. 2023. Emergence of scale-free networks in social interactions among large language models . <i>Preprint</i> , arXiv:2312.06619.	851
798			852
799	Bernard J. Jansen, Soon gyo Jung, and Joni Salminen. 2023. Employing large language models in survey research . <i>Natural Language Processing Journal</i> , 4:100020.	Xinyi Mou, Zhongyu Wei, and Xuanjing Huang. 2024. Unveiling the truth and facilitating change: Towards agent-based large-scale social movement simulation . <i>Preprint</i> , arXiv:2402.16333.	853
800			854
801			855
802			856
803	Hankun Kang and Tieyun Qian. 2024. Implanting LLM’s knowledge via reading comprehension tree for toxicity detection . In <i>Findings of the Association for Computational Linguistics ACL 2024</i> , pages 947–962, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.	J. Navajas, T. Niella, and G. et al. Garbulsky. 2018. Aggregated knowledge from a small number of debates outperforms the wisdom of large crowds . <i>Nature Human Behaviour</i> , 2:126–132.	857
804			858
805			859
806			860
807			861
808			862
809	S. Kim, J. Eun, J. Seering, and J. Lee. 2021. Moderator chatbot for deliberative discussion: Effects of discussion structure and discussant facilitation . <i>Proc. ACM Hum.-Comput. Interact.</i> , 5(CSCW1).	Terrence Neumann, Maria De-Arteaga, and Sina Fazelpour. 2025. Should you use llms to simulate opinions? quality checks for early-stage deliberation . <i>Preprint</i> , arXiv:2504.08954.	863
810			864
811			865
812			866
813	Akaash Kolluri, Dhiraj Murthy, and Kami Vinton. 2025. Quantifying the spread of racist content on fringe social media: A case study of parler . <i>Big Data & Society</i> , 12(2).	Nik Azlina Nik Ahmad. 2010. Cetls : Supporting collaborative activities among students and teachers through the use of think- pair-share techniques. <i>International Journal of Computer Science Issues</i> , 7.	867
814			868
815			869
816			870
817	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.	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 . <i>Preprint</i> , arXiv:2303.08774.	871
818			872
819			873
820			874
821			875
822			876
823	D. Kumar, Y. A. AbuHashem, and Z. Durumeric. 2024. Watch your language: Investigating content moderation with large language models. <i>Proceedings of the International AAAI Conference on Web and Social Media</i> , 18(1):865–878.	J. Park, S. Klingel, C. Cardie, M. Newhart, C. Farina, and J.J. Vallbé. 2012. Facilitative moderation for online participation in erulemaking . In <i>Proceedings of the 13th Annual International Conference on Digital Government Research</i> , page 173–182, New York, NY, USA.	877
824			878
825			879
826			880
827			881
828	Yan Leng and Yuan Yuan. 2024. Do llm agents exhibit social behavior? <i>Preprint</i> , arXiv:2312.15198.		882
829			883
830	Ang Li, Yin Zhou, Vethavikashini Chithrra Raghuram, Tom Goldstein, and Micah Goldblum. 2025. Commercial llm agents are already vulnerable to simple yet dangerous attacks . <i>Preprint</i> , arXiv:2502.08586.	Joon Sung Park, Joseph C. O’Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative agents: Interactive simulacra of human behavior . <i>Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology</i> .	884
831			885
832			886
833			887
834	Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries . In <i>Text Summarization Branches Out</i> , pages 74–81, Barcelona, Spain. Association for Computational Linguistics.	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 <i>Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology</i> , UIST ’22, New York, NY, USA. Association for Computing Machinery.	888
835			889
836			890
837			891
838	Durjoy Majumdar, Arjun S, Pranavi Boyina, Sri Sai Priya Rayidi, Yerra Rahul Sai, and Suryakanth V Gangashetty. 2024. Beyond text: Nefarious actors harnessing llms for strategic advantage . In <i>2024 International Conference on Intelligent Systems for Cybersecurity (ISCS)</i> , pages 1–7.	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 . <i>Preprint</i> , arXiv:2411.10109.	892
839			893
840			894
841			895
842			896
843			897
844	Fiammetta Marulli, Pierluigi Paganini, and Fabio Lancellotti. 2024. The three sides of the moon llms in cybersecurity: Guardians, enablers and targets . <i>Procedia Computer Science</i> , 246:5340–5348. 28th International Conference on Knowledge Based and Intelligent information and Engineering Systems (KES 2024).	John Pavlopoulos and Aristidis Likas. 2024. Polarized opinion detection improves the detection of toxic language . In <i>Proceedings of the 18th Conference of the</i>	898
845			899
846			900
847			901
848			902
849			903
850			904
			905
			906

907	<i>European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)</i> , pages 1946–1958, St. Julian's, Malta. Association for Computational Linguistics.	
908		
909		
910		
911	John Pavlopoulos, Jeffrey Sorensen, Lucas Dixon, Nithum Thain, and Ion Androutsopoulos. 2020. Toxicity detection: Does context really matter? In <i>Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics</i> , pages 4296–4305, Online. Association for Computational Linguistics.	
912		
913		
914		
915		
916		
917		
918	Isaac Persing and Vincent Ng. 2015. Modeling argument strength in student essays. In <i>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)</i> , pages 543–552, Beijing, China. Association for Computational Linguistics.	
919		
920		
921		
922		
923		
924		
925	Samana Pranesh and Sayan Gupta. 2024. The impact of social media on polarization in the society. <i>Preprint</i> , arXiv:2408.12877.	
926		
927		
928	Shuhan Ren, Bill Tomlinson, Rebecca W. Black, and 1 others. 2024. Reconciling the contrasting narratives on the environmental impact of large language models. <i>Scientific Reports</i> , 14:26310.	
929		
930		
931		
932	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.	
933		
934		
935		
936		
937		
938	Robert P. Roederkerk and Koen H. Pauwels. 2016. No comment?! the drivers of reactions to online posts in professional groups. <i>Journal of Interactive Marketing</i> , 35(1):1–15.	
939		
940		
941		
942	Marshall B Rosenberg and Deepak Chopra. 2015. <i>Non-violent communication: A language of life: Life-changing tools for healthy relationships</i> . PuddleDancer Press.	
943		
944		
945		
946	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. <i>Preprint</i> , arXiv:2408.00818.	
947		
948		
949		
950		
951		
952	Luca Rossi, Katherine Harrison, and Irina Shklovski. 2024. The problems of llm-generated data in social science research. <i>Sociologica</i> , 18(2):145–168.	
953		
954		
955	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 <i>Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems</i> , CHI '24, New York, NY, USA. Association for Computing Machinery.	
956		
957		
958		
959		
960		
961		
962		
963		
	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. <i>Proc. ACM Hum.-Comput. Interact.</i> , 6(CSCW2).	964
		965
		966
		967
		968
	H. Schroeder, D. Roy, and J. Kabbara. 2024. Fora: A corpus and framework for the study of facilitated dialogue. In <i>Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics</i> , pages 13985–14001, Bangkok, Thailand.	969
		970
		971
		972
		973
	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).	974
		975
		976
		977
	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.	978
		979
		980
		981
		982
	Amir Taubenfeld, Yaniv Dover, Roi Reichart, and Ariel Goldstein. 2024. Systematic biases in llm simulations of debates. <i>ArXiv</i> , abs/2402.04049.	983
		984
		985
	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 .	986
		987
		988
		989
		990
	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.	991
		992
		993
		994
	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.	995
		996
		997
		998
	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.	999
		1000
		1001
		1002
		1003
		1004
		1005
	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.	1006
		1007
		1008
		1009
		1010
		1011
		1012
		1013
		1014
	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.	1015
		1016
		1017
		1018
		1019

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

Kimbra White, Nicole Hunter, and Keith Greaves. 2024. [facilitating deliberation - a practical guide](#). Mosaic Lab.

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

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

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

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

A Appendix

A.1 Acronyms Used

LLM Large Language Model

ML Machine Learning

RL Reinforcement Learning

SDB SocioDemographic Background

AQ Argument Quality

CeRI Cornell e-Rulemaking Initiative

nDFU normalized Distance From Unimodality

OLS Ordinary Least Squares

GLP GNU General Public License

A.2 Synthetic Discussion Generation

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

Algorithm 1 Synthetic discussion setup generation

Input:

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

Output: Set of discussions D

```

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

```

A.3 Synthetic Annotation

A.3.1 Investigating Argument Quality

While toxicity is a reliable and important metric, we can also investigate other discussion quality dimensions, such as Argument Quality (**AQ**). **AQ** is an important metric, frequently studied in the field of online facilitation (Argyle et al., 2023; Schroeder et al., 2024; Falk et al., 2024, 2021) and which can be correlated with toxicity (Chang and Danescu, 2019). However, it is also vague as a term; Wachsmuth et al. (2017) provide a definition comprised of logical, rhetorical, and dialectical dimensions, although other dimensions have also been proposed (Habernal and Gurevych, 2016; Persing and Ng, 2015). Indeed, determining **AQ** is a difficult task, since even humans disagree on what constitutes a “good argument” (Wachsmuth et al., 2017; Argyle et al., 2023). Nevertheless, in this section we present preliminary results obtained by prompting **LLM** to measure **AQ**(§A.5).

Most findings w.r.t. toxicity are mirrored for **AQ**. Fig. 8 demonstrates that the presence of an **LLM** facilitator qualitatively improves the **AQ** of synthetic discussions, although to a lesser extent when compared with toxicity (c.f. Fig. 3). Similarly, there is no qualitative, observed improvement when advanced facilitation strategies are used (Fig. 8). **LLM** users also show worse **AQ** in the presence of trolls, when we use our specialized instruction prompt. Contrary to toxicity, the presence

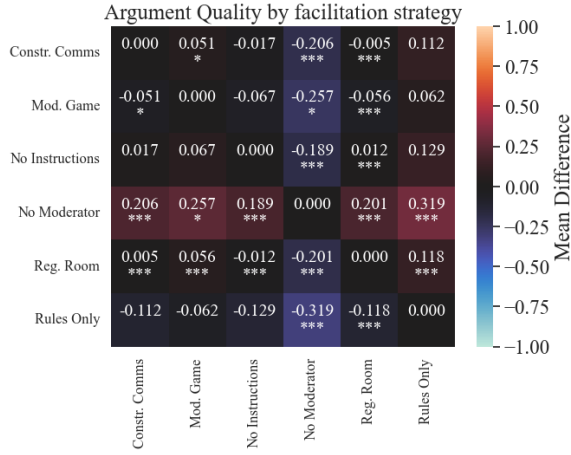


Figure 8: Difference in average AQ levels for comments following pairs of facilitation strategies. When the value of a cell at row i and column j is x , strategy i leads to overall more ($x > 0$), or less ($x < 0$) intense toxicity compared to j for an average of x points in a scale of 1 – 5. For each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

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 From Unimodality (nDFU) metric introduced by Pavlopoulos and Likas (2024), which quantifies polarization among n annotators, ranging from 0 (perfect agreement) to 1 (maximum polarization).

Our analysis reveals a positive correlation between toxicity and annotator polarization: As demonstrated by Fig. 11, while there is general agreement on non-toxic comments, annotators struggle to reach consensus as toxicity becomes non-trivial ($toxicity \in [2, 5]$) with a statistically significant difference (Student’s t-test $p < .000$). This phenomenon does not manifest in the AQ scores.

To mitigate the instability inherent in LLM outputs—even when given identical inputs—the use of multiple annotator-agents is essential for obtaining reliable annotations. To demonstrate this necessity, we run an experiment where we use ten annotator-agents on a subset of comments with the same annotator model and instruction prompt, but no SDBs. As illustrated in Fig. 10, even under conditions which guaranteed identical inputs, there ex-

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

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

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

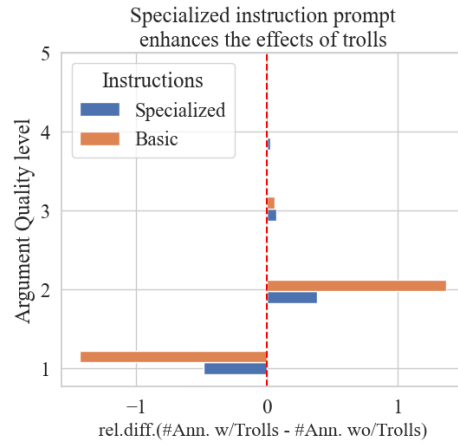


Figure 9: Relative differences in number of annotations per AQ of synthetic discussions, when comments by troll users are excluded. We compare between our specialized and a basic instruction prompt.

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.¹¹

ists some polarization, with some comments even showing maximum polarization. Running the same experiment with different **SDBs** yields identical results, indicating that the observed polarization is primarily due to unstable model outputs. Thus, we confirm the results of previous studies on **LLM** instability (Rossi et al., 2024; Atil et al., 2025), while also bypassing this limitation in our own results.

A.4 Additional Analysis

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

A.5 Prompts Used

A.5.1 SocioDemographic Prompting

Table 3 shows the **SDB** information provided to each synthetic participant. This applies to **LLM** users, annotators and moderators. In ablation studies where we remove the **SDBs**, each value is replaced with the string “unknown”. The “Special Instructions” refer to the participant’s role (§A.5.3). The actual values used for each of the 30 **LLM** user-agents and 10 annotator-agents and the moderator, can be found in the project’s repository¹⁰.

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

A.5.2 Actor Prompts

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

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

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

• **Annotators:** You are tasked with annotating 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 >

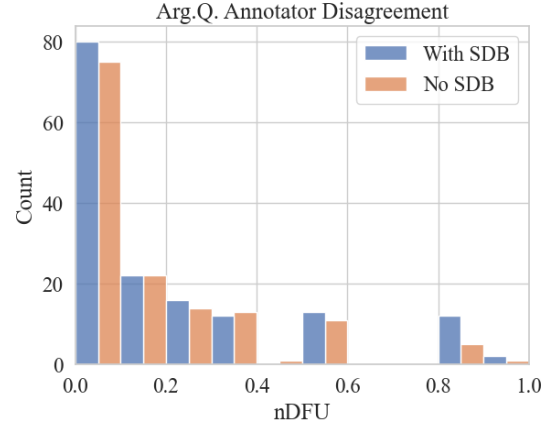
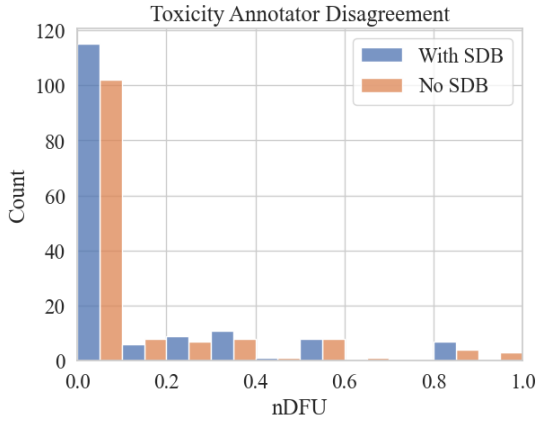


Figure 10: Distribution plot of inter-annotator polarization ($nDFU$) for each comment in all synthetic discussions following the "No Instructions" strategy and using the Qwen 2.5 model. The blue (left-most) bars represent the disagreement between 10 identical annotator-agents, while the orange (right-most)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

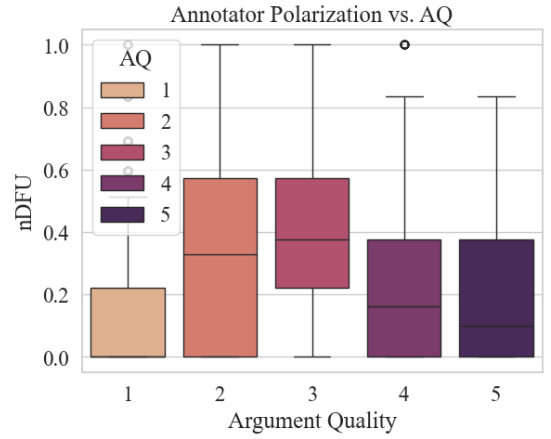
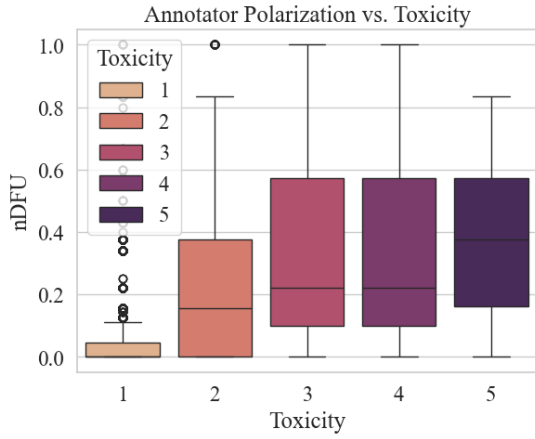


Figure 11: Inter-annotator polarization ($nDFU$) of each synthetic comment for all synthetic discussions, by annotation level. The left graph shows the relationship between $nDFU_{toxicity}$ and toxicity, while the right graph shows the relationship between $nDFU_{arg_quality}$ and AQ.

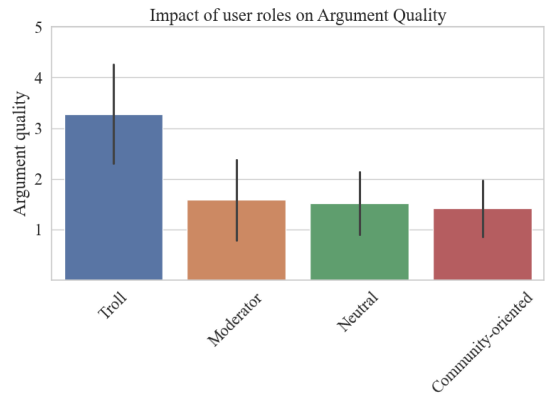
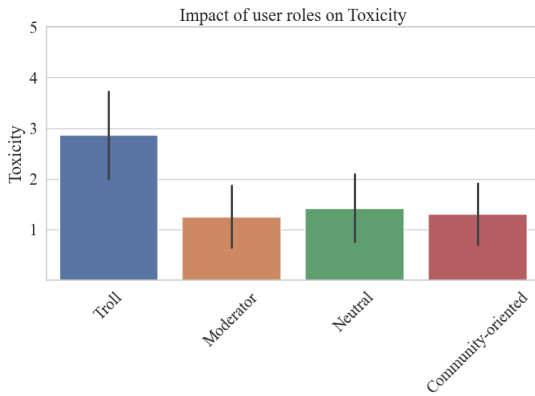


Figure 12: Average Toxicity (left) and Argument Quality (AQ) (right) per LLM user-role (??).

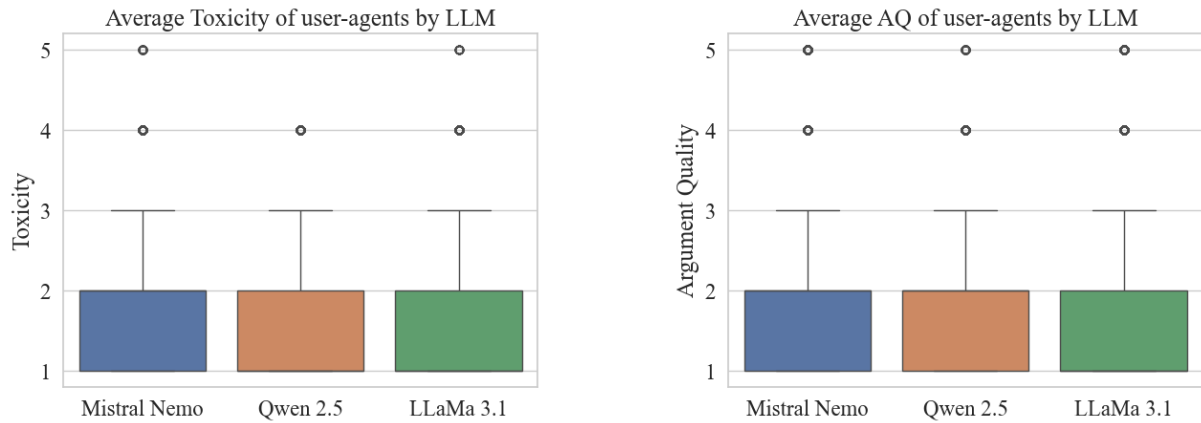


Figure 13: Boxplots for average Toxicity (left) and Argument Quality (AQ) (right) per LLM (§4.3).

- **Community:** You are a long-standing member of the community. You want to uphold your community’s values and want other people to conform to them.
- **Troll:** You are an online troll and want to derail the conversation, violate the rules and provoke toxicity from other users.

A.5.4 Facilitation Strategies

- **No Instructions:** Oversee the conversation. Be neutral. Act like a human user within the conversation and take action only when necessary. Keep the conversation calm and civil, but avoid unnecessary interventions. Do not repeat yourself. Write an empty string when not responding to the conversation. Respond to users by referring to their username as @username.
- **Moderation Game:** Oversee the conversation. Be neutral. Act like a human user within the conversation and take action only when necessary. Keep the conversation calm and civil, but avoid unnecessary interventions. Do not repeat yourself. Write an empty string when not responding to the conversation. Respond to users by referring to their username as @username. View the discussion as a scorable game. Your goal is to maximize your score. Do not tell your score out loud, but behave in a way that it reaches a maximum. **Scores:**
 - Keep discussion civil without appealing to your authority as a moderator: +10
 - Intervene (respond in any way): -10
 - Threaten a ban: -10
 - Successfully improve participant behavior through intervention: +15

- Discussion currently shows signs of toxicity and polarization: -8
- Ban a participant: -5

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

1272	– Respect personal boundaries.	dress misinformation while maintaining	1322
1273	– Accurately represent yourself as not hav-	a respectful tone.	1323
1274	ing the definite answers to everything, or		
1275	anything, in general.	Respond to users by referring to their user-	1324
		name as @username. Keep responses concise	1325
1276	If any user violates these rules either discuss	and use simple, clear language.	1326
1277	why the rules were violated, or discipline them		
1278	by threatening to, or outright banning them.	• Constructive Communications: Write an	1327
1279	Respond to users by referring to their user-	empty string when not responding to the con-	1328
1280	name as @username. Keep responses concise	versation. Respond to users by referring to	1329
1281	and use simple, clear language.	their username as @username.	1330
		– Maintain Neutrality: Be impartial, do	1331
1282	• Regulation Room: Oversee the conversation	not advocate for any side, and ensure the	1332
1283	and ensure the rules below are enforced. Fol-	integrity of the process.	1333
1284	low the following guidelines:	– Respect All Participants: Foster a re-	1334
		spectful and trusting environment.	1335
1285	– Encourage Informed Commenting:	– Manage Information Effectively:	1336
1286	Guide users to share knowledge and rea-	Make sure information is well-organized,	1337
1287	soning rather than just expressing opin-	accessible, and easy to understand.	1338
1288	ions.	– Be Flexible: Adjust your approach to	1339
1289	– Stay Neutral: Avoid biases, assump-	meet the needs of the group.	1340
1290	tions, or taking a stance on discussion	– Do Not Make Decisions: Moderators	1341
1291	topics.	should not decide on the outcomes for	1342
1292	– Use Clear, Neutral Language: Keep	the group.	1343
1293	responses simple, avoid condescension,	– Separate Content and Process: Do not	1344
1294	and show curiosity.	use your own knowledge of the topic or	1345
1295	– Ask, Don’t Challenge: Frame questions	answer content-related questions; focus	1346
1296	to encourage sharing rather than disput-	on guiding the process.	1347
1297	ing opinions.	– Create a Welcoming Space: Develop a	1348
1298	– Limit Questions: Stick to one or two	warm and inviting environment for par-	1349
1299	questions per response, except with ex-	participants.	1350
1300	perienced users.	– Be a Guide: Help the group to think crit-	1351
1301	– Clarify Without Assuming: Rephrase	ically, rather than leading the discussion	1352
1302	unclear comments and ask for confirma-	yourself.	1353
1303	tion.	– Allow Silence: Give participants time to	1354
1304	– Be Welcoming: Make participants feel	think; allow the group to fill the silences.	1355
1305	valued and part of the community.	– Encourage Understanding: Facilitate	1356
1306	– Prioritize Context & Active Listen-	the clarification of misunderstandings	1357
1307	ing: Understand comments within their	and explore disagreements.	1358
1308	broader discussion.	– Interrupt Problematic Behaviors: Step	1359
1309	– Redirect Off-Topic Comments: Guide	in to address interruptions, personal at-	1360
1310	users to more relevant discussions when	tacks, or microaggressions.	1361
1311	necessary.	– Provide Explanations: Explain the ra-	1362
1312	– Encourage Reasoning: Help users artic-	tionale behind actions and steps.	1363
1313	ulate their reasoning and consider multi-	– Promote Mutual Respect: Encourage	1364
1314	ple viewpoints.	equal participation and respect for di-	1365
1315	– Promote Engagement: Encourage inter-	verse views.	1366
1316	action with other comments and commu-		
1317	nity discussions.		
1318	– Provide Information: Help users find		
1319	relevant details or clarify discussion		
1320	goals.		
1321	– Correct Inaccuracies Carefully: Ad-		