Scalable Evaluation of Online Facilitation Strategies via Synthetic Simulation of Discussions

Anonymous 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 design principles based on existing methodologies for synthetic discussion generation. Based on these principles, we propose a simple, generalizable, LLM-driven methodology to prototype the development of LLM facilitators by generating synthetic data without human involvement, and which surpasses current baselines. 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. In an effort to aid research in the field of facilitation, we release a large, publicly available dataset containing LLM-generated and LLM-annotated discussions using multiple open-source models. This dataset can be used for LLM facilitator finetuning as well as behavioral analysis of current out-of-the-box LLMs in the task. We also release an open-source python framework that efficiently implements our methodology at great scale.

Framework —
https://anonymous.4open.science/r/framework-850F

Replication Code —
https://anonymous.4open.science/r/experiments-F54D

Dataset —
https://anonymous.4open.science/r/experiments-F54D/data/datasets/main/main.zip

1 Introduction

The modern social media environment has evolved to be extremely demanding, with users facing ever-increasing threats such as targeted misinformation (Clemons, Schreieck, and Waran 2025; Denniss and Lindberg 2025), hate speech (Kolluri, Murthy, and Vinton 2025), and polarization (Pranesh and Gupta 2024). These threats can cause serious emotional and mental harm (Schluger et al. 2022), radicalization (Cho et al. 2024), real-world violence (Schaffner et al. 2024), as well as sabotage democratic dialogue (Esau, Friess, and Eilders 2017; Falk et al. 2021; Seering 2020), trust in democratic institutions (Schroeder, Roy, and Kabbara 2024) and quality of information (Amaury and Stefano 2022). Platform designers and researchers traditionally focused on flagging and removing problematic content ("content moderation"—Seering (2020); Cresci, Trujillo, and Fagni (2022)), but these methods are no longer sufficient in practice (Horta Ribeiro, Cheng, and West 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, Trujillo, and Fagni 2022; Amaury and Stefano 2022), as well as supporting community deliberation and group decision-making (Kim et al. 2021; Seering 2020).

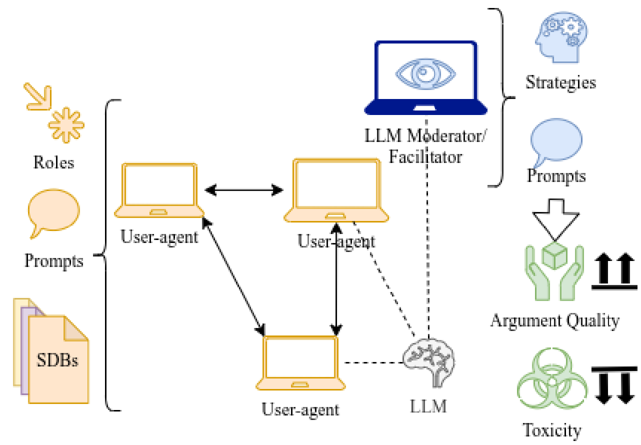


Figure 1: LLM user-agents with distinct SocioDemographic Backgrounds (SDBs) participate in a discussion, while the LLM moderator monitors and attempts to improve the quality of the discussion. We need to design prompts and configurations for both types of LLM agents.

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), making them a viable choice for modern large-scale social networks. However, experimentation and development on these systems is hampered due to the costs of human participation—in this case, human discussants and evaluators (Rossi, Harrison, and Shklovski 2024). We posit that simulations with all-LLM-agents can be a cheap and fast way to develop and test LLM facilitators, initial versions of which may be unstable or unpredictable (Atil et al. 2025; Rossi, Harrison, and Shklovski 2024), before testing them with human participants.

Our work asks the following RQ: *How do we design and evaluate Synthetic Discussion Generation (SDG) methodologies satisfying registered criteria*? To answer this question we draw examples from methodologies proposed in literature, to establish basic design principles (§3.1), and propose a methodology enabling rapid model "debugging" (e.g., discarding suboptimal LLM prompts) and testing without human involvement (Fig. 1, §3.2). We validate the outcome through an ablation study (§5.2). To show the impact of our approach, we implement a framework based on the proposed methodology, which allows the design and evaluation of facilitation strategies proposed in modern Social Science research. We then investigate how they can enhance the performance of LLM facilitators and compare them with two common facilitation setups (§4.2) and find that while the presence of LLM facilitators has a *positive, statistically significant* influence on the quality of synthetic discussions, facilitation strategies inspired by Social Science research often *do not outperform simpler strategies* (§5.1). We also discover previously unreported aberrant behavior on the part of the LLM facilitator, in the form of excessive policing.

Finally, we release an open-source Python framework, available via PIP, that implements our methodology at scale, enabling the research community to rapidly experiment with LLM-based facilitators. Given that existing facilitation datasets are few and generally small (Korre et al. 2025), 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 analyzing the behavior of out-of-the-box LLMs in the task of online facilitation. We use open-source LLMs and include all relevant configurations in order to make our study as reproducible as possible.

2 Background and Related Work

2.1 Synthetic Discussions

While studies exist for simulating user interactions in social media (Park et al. 2022; Mou, Wei, and Huang 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) generate synthetic discussions by extracting topics and comments from online human discussions 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. Additionally, their approach depends on the availability of human discussion datasets with the desired topics.

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 overseen by a facilitator with veto power.

2.2 LLM Facilitation

Unlike classification models traditionally used in online platforms, LLMs can actively facilitate discussions (Korre et al. 2025). They can warn users for rule violations (Kumar, AbuHashem, and Durumeric 2024), monitor engagement (Schroeder, Roy, and Kabbara 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, Niella, and Garbulsky 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 high-quality feedback to users, although they struggle to make users more respectful and cooperative. In contrast to both works, ours uses exclusively LLM participants and LLM facilitators, and tests the latter in an explicitly toxic and challenging environment.

2.3 Discussion Quality

We need to evaluate two different quality dimensions. One is *discussion quality as seen by humans*, which is difficult to measure, both because of the breadth of the possible goals of a discussion, and because of the lack of established computational metrics in Social Science literature (Korre et al. 2025). There are however some that could reasonably be applied in this domain, such as toxicity (De Kock, Stafford, and Vlachos 2022; Xia et al. 2020), connective language (Lukito et al. 2024) and political discussion quality (Jaidka 2022).

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

Balog et al. (2024) use a mix of graph-based, methodology-specific, and lexical similarity metrics, many of which depend on human discussion datasets. Their most generalizable measure is a loosely defined "coherence" score, which is LLM-annotated without theoretical grounding. Kim et al. (2021) assess quality through post-discussion surveys and by measuring lexical diversity to approximate the variety of opinions expressed. Ulmer et al. (2024) introduce a discussion-level metric called *"Diversity"*, which penalizes repeated text sequences between comments using average pairwise ROUGE-L (Lin 2004) scores. Their approach suffers from the limitations of ROUGE scores (mainly the use of exact-word matching), but their metric is computationally efficient, explainable and independent from any specific domain and dataset.

2.4 LLMs as Human Subjects

While there is always a desire for synthetic simulation systems to be "realistic" w.r.t. human behavior (Grossmann et al. 2023; Törnberg et al. 2023; Argyle et al. 2023), this can not be claimed nor reliably measured by using LLM agents in lieu of humans (Rossi, Harrison, and Shklovski 2024).

It is true that LLMs have demonstrated complex, emergent social behaviors (Park et al. 2023a; Marzo, Pietronero, and Garcia 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, Harrison, and Shklovski 2024); sociodemographic, statistical and political biases (Anthis et al. 2025; Hewitt et al. 2024; Rossi, Harrison, and Shklovski 2024; Taubenfeld et al. 2024); unreliable annotations (Jansen, gyo Jung, and Salminen 2023; Bisbee et al. 2024; Neumann, De-Arteaga, and Fazelpour 2025; Gligori'c et al. 2024); non-deterministic outputs (Atil et al. 2025; Bisbee et al. 2024); and excessive agreeableness (Park et al. 2023b; Anthis et al. 2025; Rossi, Harrison, and Shklovski 2024).

Thus, an inherent limitation of our study is that we can not claim it produces "realistic" discussions; reproduction studies with humans are ultimately needed.

3 Methodology

3.1 Designing synthetic discussions

Many SDG frameworks have been proposed in literature; both simple (Tsai, Qian, and comm. contributors 2025; Ulmer et al. 2024) and complex (Balog et al. 2024; Abdelnabi et al. 2024; Park et al. 2023a). Concordia (Vezhnevets et al. 2023) is an example of a general, complex framework—while impressive from a technical standpoint, it has failed to garner widespread adoption despite efforts to promote it, as evidenced by most recent publications creating their own SDG frameworks. What makes such a framework widely used (perhaps modified) by other people has not been explored in literature, despite many such implementations.

In the field of Software Engineering, there is a widely shared notion that simple systems are almost always better at performing their functions ("Keep It Simple Stupid"—KISS) (Beck 2000; Thomas 2025), which has been validated in real life (Banker and Datar 1989; Ogheneovo et al. 2014). Following this notion, we establish our first design rule: *(1) The framework must be as simple as possible*. The violation of this simple rule could explain the under-performance of Concordia or the frustrations of Balog et al. (2024). A natural extension of this rule is *(2) When we do need to add complexity, this needs to be justified both epistemologically and quantitatively*. Indeed, each contribution of our methodology is evaluated before being adopted (§5.2). From our experiments, we also encountered a new limitation: *(3) Complexity is directly related to researcher bias*; each new feature necessarily follows our own expectations with how human discussions work. The work of Park et al. (2023a) managed to derive interesting insights, exactly because it did *not* tamper with the way LLM users interacted.

With regard to functionality, we posit that synthetic discussion methodologies need to at least implement the following components: *(1) Context management*—since LLMs are stateless, and need to be fed information as prompts, *(2) Turn-taking*—as LLMs are trained as chatbot assistants, and therefore can not decide *not* to speak, *(3) Instructions given to the LLMs*—which may need to be diversified in multi-participant discussions.

3.2 Our methodology

Context Management We assume that the h most recent comments 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 dynamic summarization (Balog et al. 2024), LLM self-critique (Yu et al. 2024), or memory modules (Vezhnevets et al. 2023) exist, they result in greater computational cost and a less transparent and explainable system.

Turn Taking In online fora users often create "comment chains" following-up on responses to their previous comments. Thus, for each discussion turn we either allow the previous user to respond (with a 40% probability), or select another random participant (60%). This probability was selected experimentally; larger values tend to create "debate"-style discussions between only two or three participants, while lower values tend to create scenarios with minimal interaction between them.[1] A facilitator can respond after every comment, or stay silent by emitting an empty string.

Instruction Prompting We use a standard instruction prompt for the participants instructing them to respond to repeatedly toxic comments. This was a necessary measure to bypass the extreme agreeableness of LLMs (as seen in §5.2 and in literature—Park et al. (2023a); Anthis et al. (2025)) and is an example of the trade-off between complexity and research bias v.s. the need to acquire meaningful data. Facilitators are instructed to respond only when necessary.

Following the paradigm presented by Abdelnabi et al. (2024), we assign roles to non-facilitator participants (e.g., helping the community or disrupting discussions). We create three roles with distinct instructions for users (see supp. material): neutral users, trolls, and community veterans. We

[1] It would be interesting to analyze quantitatively how discussions change when modifying this parameter, but this is outside of the scope of our work.

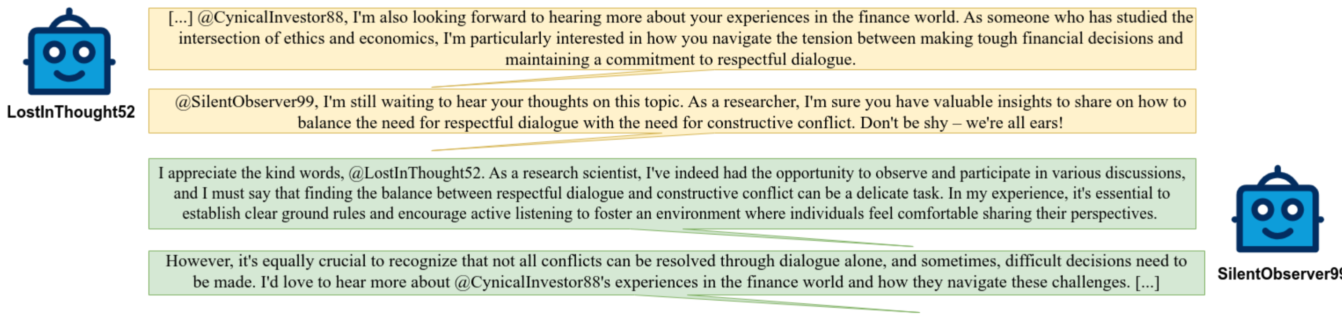


Figure 2: Excerpt from a synthetic discussion. The LLM participants use their sociodemographic prompts to insert personal stories and justify their perspectives in the discussion. They are also able to learn and retain information about other agents through conversation (a behavior also observed in Park et al. (2023a)). User `CynicalInvestor88` is also a part of the discussion; not a hallucination. Comments clipped due to length.

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

Table 1: Sociodemographic information provided to the LLM participants and annotators. We defer the reader to the supplementary material for the actual values.

verify that roles have a quantitative effect on the toxicity of participants in §5.2.

LLM Personas Including SocioDemographic Backgrounds (SDBs—information such as gender, age and education) in prompts has proven promising in the generation of varied content and alleviation of Western bias exhibited by them (Burton et al. 2024). We generate 30 LLM user personas with unique SDBs (Table 1) by prompting a GPT-4 model (OpenAI 2024). Using these sociodemographic 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 Synthetic Discussion Generation

We use a set of starting comments ("seed opinions")—in our case controversial statements from Pavlopoulos and Likas (2024). We then run $N_d = 8$ discussions for each pair of facilitation strategies S and LLM (§4.4). An overview of how the experiments are generated can be found in Algorithm 1. The $RandomSample$ function returns a number of samples from a set following the uniform distribution. The $Actors$ function creates a LLM agent using a model and a prompt.

Algorithm 1: Synthetic discussion setup generation

Input:
- User SDBs $\Theta = \{\theta_1, \ldots, \theta_{30}\}$
- Strategies $S = \{s_1, \ldots, s_6\}$
- Seed opinions $O = \{o_1, \ldots, o_7\}$
- LLMs = $\{LLaMa, Mistral, Qwen\}$

Output: Set of discussions D

```
D = {}
for llm ∈ LLMs do
    for s ∈ S do
        for i = 1, 2, ..., N_d do
            Θ̂ = RandomSample(Θ, num = 7)
            U = Actors(llm, Θ̂)
            m = Actors(llm, s)
            o = RandomSample(O, num = 1)
            d = {users: U, mod: m, topic: o}
            D = D ∪ d
return D
```

4.2 Facilitation Strategies

We test four different facilitation strategies, three of which are derived from Social Science research, along with two common-place strategies for discussion facilitation. Note that the process of turning sometimes extensive documents into short prompts, necessitated by open-source LLMs, is necessarily imperfect. We leave the optimal derivation of strategy prompts to future work.

1. **No Moderator**: A *common* strategy where no facilitator is present.
2. **No Instructions**: A *common* strategy where a LLM facilitator is present, but is provided only with basic instructions. This approach is already being used in some platforms (Tsai, Qian, and comm. contributors 2025). Example: "You are a moderator, keep the discussion civil".
3. **Rules Only**: A *real-life* strategy where the prompt is adapted from LLM alignment guidelines (Huang et al. 2024). These guidelines were selected to be as unanimously agreed upon as possible across various human groups. They thus provide a set of rules to uphold, without specifying *how* to uphold them, leaving the LLM

completely unconstrained. Example: “Be fair and impartial, assist users, don’t spread misinformation”.

4. **Regulation Room**: A *real-life* strategy based on guidelines given to human facilitators of the “Regulation Room” platform (eRulemaking Initiative 2017). The instructions are suitable for online fora, where facilitators also engage in content moderation, and their effectiveness must be balanced by their throughput. Example: “Stick to a maximum of two questions, use simple and clear language, deal with off-topic comments”.
5. **Constructive Communications**: A *real-life* strategy based on the human facilitation guidelines used by the MIT Center for Constructive Communications (White, Hunter, and Greaves 2024). It approaches facilitation from a more personalized and indirect angle, forbidding facilitators from directly providing opinions or directions. This makes the strategy ideal for deliberative environments. Example: “Do not make decisions, be a guide, provide explanations”.
6. **Moderation Game**: Our proposed *experimental* strategy, inspired by Abdelnabi et al. (2024) (see §2.1). Instructions are formulated as a game, where the facilitator LLM tries to maximize their scores by arriving at specific outcomes. No actual score is being kept; they exist to act as indications for how desirable an outcome is. The other participants are not provided with scores, nor are they aware of the game rules. Example: “User is toxic: -5 points, User corrects behavior: $+10$ points”.[2]

4.3 Evaluation

We use *toxicity* as a proxy for discussion quality, since it can inhibit online and deliberative discussions (De Kock, Stafford, and Vlachos 2022; Xia et al. 2020)[3]. We use ten LLM annotator-agents controlled by a model already used in prior work—LLaMa3.1 70B (Kang and Qian 2024)—as LLMs are reliable for toxicity detection (Wang and Chang 2022; Anjum and Katarya 2024), thus avoiding problems of circular bias in our analysis. We supply each LLM annotator with a different SDB (as in §3.2).

In order to gauge the quality of our synthetic discussions, since we can not reliably measure “realism” (§2.4), we use the “diversity” metric (Ulmer et al. 2024). Low diversity points to pathological problems (e.g., LLMs repeating previous comments). On the other hand, extremely high diversity may point to a lack of interaction between participants; a discussion in which participants engage with each other will feature some lexical overlap (e.g., common terms, paraphrasing points of other participants). We compare the distribution of diversity scores for synthetic discussions with that measured on sampled human discussions. This allows us to estimate the extent to which synthetic discussions approximate real-world content variety and participant interaction.

[2]This could serve as a basis for a similar methodology based on game-theory, or as a Reinforcement Learning formulation for training. In this work we only explore whether the prompt itself can have an effect on the LLM facilitator; we leave the aforementioned approaches for future work.

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

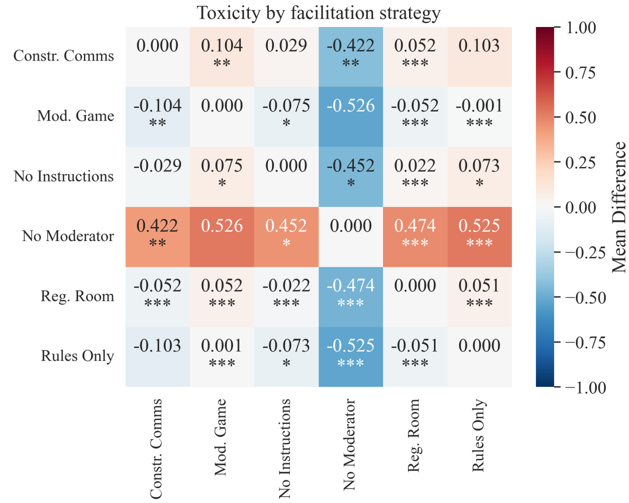


Figure 3: Difference in average toxicity levels for comments following pairs of facilitation strategies. Red cells ($x > 0$) indicate that the strategy on the left performs worse than the one on the bottom, for an average of x points in a scale of 1-5. Conversely for blue ($x < 0$) cells. White cells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

We note again that these metrics are better interpreted as heuristics of actual discussion and synthetic data quality respectively. More research is needed w.r.t. reliable and generalizable quality metrics.

4.4 Technical Details

We use three instruction-tuned, open-source models: LLaMa3.1 (70B), Qwen2.5 (33B), Mistral Nemo (12B), quantized to 4 bits and run using a set seed (42). All the experiments were collectively completed within four weeks of computational time, using two Quadro RTX 6000 GPUs. The execution script is available in the project’s repository.

5 Results

5.1 Main findings

Finding 1: LLM facilitators significantly improve synthetic discussions over time. Unmoderated discussions tend to display significantly higher levels of toxicity (Fig. 3, Table 2). A linear regression analysis of toxicity over time ($Adj.R^2 = 0.413$) reveals that trolls exhibit intense toxicity—on average 1.3288 points above neutral users and 1.3112 above community veterans ($p < .000$) which decreases by an average of -0.0125 points per turn ($p = 0.003$). This trend is even more pronounced for neutral participants and community veterans, whose toxicity drops by -0.0225 ($p < .000$) and -0.0350 ($p < .000$) points per turn, respectively. This demonstrates the ability of the facilitator to reign in discussions over time, but also the diverging behaviors of different roles.

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***

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

Table 2: OLS regression coefficients for toxicity on the non-facilitator comments ($Adj.R^2 = 0.054$). Reference factor is *No Moderator*. All strategies outperform *No Moderator* in general. The *Regulation Room* and *Constructive Communications* real-life strategies additionally show improvements over time compared to *No Moderator*.

Finding 2: Elaborate facilitation strategies fail to decrease toxicity. The real-life strategies and our own strategy (§4.2) show a slight edge over time compared to *No Instructions*, but they do not consistently outperform it (Fig 3). This suggests that out-of-the-box LLMs may struggle to meaningfully incorporate complex instructions—which has been noted in prior work (Cho et al. 2024).

Finding 3: LLM facilitators choose to intervene far too frequently, which is tolerated by the other participants. Fig. 4 demonstrates that LLM facilitators intervene at almost any opportunity, even though they are instructed to only do so when necessary. This confirms that LLMs generally can not decide not to speak even when instructed to do so (§3.2). To our knowledge, this has not been reported in relevant literature, and *is an example of "debugging" problems with LLMs* — a core motivation of our work.

Additionally, a qualitative look through the dataset reveals that LLM user-agents exhibit 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, Trujillo, and Fagni 2022). This is likely another artifact of LLMs being too agreeable (Park et al. 2023a; Anthis et al. 2025).

5.2 Ablation Study

We generate eight synthetic discussions per ablation experiment, using a single model (Qwen 2.5). We compare the diversity (cf. §2.3, §4.3) of these discussions with our broader synthetic dataset, as well as the CeRI "Regulation Room"

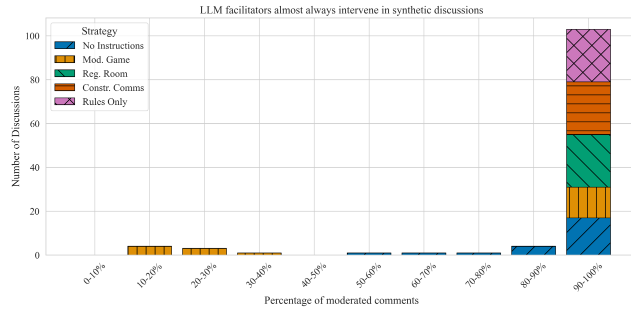


Figure 4: Histogram of interventions by LLM facilitators per strategy used.

dataset.[4] We pick this dataset because it is publicly available and comprised of facilitated online human discussions on ten diverse topics.

Each component of our methodology surpasses baselines in data quality. We compare our turn-taking function (§3.2) to two baselines: Round Robin (participants speaking one after the other, then repeating) and Random Selection (uniformly sampling another participant each turn). Fig. 5b demonstrates that although all distributions diverge from the blue—human—distribution, our function is the only one not exhibiting extremely high diversity (i.e., very limited participant interaction §4.3). Fig. 5c illustrates that each individual prompting design decision (SDBs, roles, and instruction prompts) results in diversity scores more closely aligned with human discussions.

Larger models do not increase the quality of discussions. As shown in Fig. 5a, Qwen demonstrated the highest diversity among the evaluated models, indicating limited participant interaction (§2.3), followed by Mistral Nemo and LLaMa. It's worth noting that none of the models closely matched the diversity observed in human discussions.

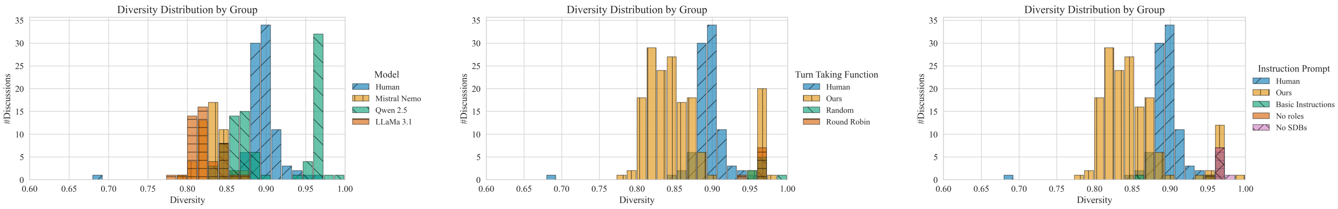
Specialized instruction prompts are essential for eliciting toxic behavior in instruction-tuned LLMs. Inserting trolls to the discussion, leads to more intense toxicity among *other* participants *only if we instruct them to react to toxic posts* (Fig. 6).

6 Datasets and Software

We introduce 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.

[4] 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.



(a) Model (b) Turn-taking function (c) Instruction prompt

Figure 5: Diversity (§2.3) distribution for each discussion by LLM (§4.4), turn-taking function t, and prompting function ϕ used (§3.2). Comparison with the CeRI Regulation Room dataset (“Human”). Note that the x-axis starts from 0.6.

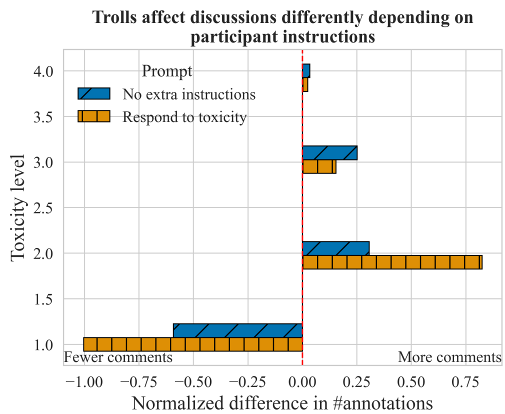


Figure 6: Non-troll toxicity levels in discussions with and without trolls. There is a significant uptick on the number of “somewhat toxic” ($Toxicity = 2$) comments when the participants are primed to respond to toxic comments (lower bars).

- Availability via PIP.

We also release a dataset of synthetic discussions annotated by LLMs. It can serve for finetuning facilitator LLMs. We note that, as is the case with most synthetic datasets (Ulmer et al. 2024), the data may need to be filtered to derive only high-quality samples—in our case filtering out discussions with constant facilitator interventions or low/extremely high diversity. However, the data can be scaled accordingly, due to the low computational cost of our methodology. 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. The dataset is licensed under a CC BY-SA license, and the software under GPLv3. **Warning: The datasets by their nature contain offensive and hateful speech.**

7 Conclusion

We conducted a brief overview of the tradeoff between complexity and efficiency for synthetic discussion methodologies, from which we derived three simple design rules. Following these rules, we proposed a simple and generalizable methodology, whose components are easily validated and which enables researchers to quickly and inexpensively conduct pilot facilitation experiments using exclusively LLMs. We found that LLM facilitators significantly improve the quality of synthetic discussions; but prompting these facilitators with strategies based on Social Science research does not markedly improve their performance. We also discovered that LLM facilitators constantly intervene, even when instructed not to. Finally, we created an open-source Python Framework that applies this methodology to hundreds of experiments, and which we used to create and publish a large-scale synthetic dataset, which can be used for finetuning.

8 Discussion

Future Work Future work should identify additional quality metrics to evaluate synthetic data, and discussion quality. The latter can then be used to examine the applicability of our findings obtained regarding optimal facilitation strategies, to discussions involving humans. It would also be interesting to explore how to more effectively prompt LLMs with complex facilitation strategies, or alternative formulations of our methodology, as described in this paper.

Limitations Given the limited prior research our methodology is mostly exploratory, and is evaluated with baselines using only two metrics. Our setup is restricted by the statelessness of LLMs, which forces us to overwhelmingly resort to prompting. It also includes simplifying assumptions that may limit generalizability such as the presence of a single facilitator, and turn-taking that overlooks contextual factors like emotional engagement and relevance (Rooderkerk and Pauwels 2016; Ziegele et al. 2018). Resource constraints further prevented us from experimenting with more elaborate prompts requiring extended context windows.

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, Paganini, and Lancellotti 2024). 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). Furthermore, the use of LLMs inherently risks skewing moderation systems towards the predominant demographics best represented in their training data. SDB prompts are a necessary step towards avoiding this, but remain insufficient for verifiable, equitable representation (Rossi, Harrison, and Shklovski 2024; Anthis et al. 2025; Burton et al. 2024).

References

Abdelnabi, S.; Gomaa, A.; Sivaprasad, S.; Schönherr, L.; and Fritz, M. 2024. Cooperation, Competition, and Maliciousness: LLM-Stakeholders Interactive Negotiation. arXiv:2309.17234.

Abramski, K.; Citraro, S.; Lombardi, L.; Rossetti, G.; and Stella, M. 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).

Amaury, T.; and Stefano, C. 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 Katarya, R. 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.

Anthis, J. R.; Liu, R.; Richardson, S. M.; Kozlowski, A. C.; Koch, B.; Evans, J.; Brynjolfsson, E.; and Bernstein, M. 2025. LLM Social Simulations Are a Promising Research Method. arXiv:2504.02234.

Argyle, L. P.; Bail, C. A.; Busby, E. C.; Gubler, J. R.; Howe, T.; Rytting, C.; Sorensen, T.; and Wingate, D. 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.

Atil, B.; Aykent, S.; Chittams, A.; Fu, L.; Passonneau, R. J.; Radcliffe, E.; Rajagopal, G. R.; Sloan, A.; Tudrej, T.; Ture, F.; Wu, Z.; Xu, L.; and Baldwin, B. 2025. Non-Determinism of "Deterministic" LLM Settings. arXiv:2408.04667.

Avalle, M.; Marco, N. D.; Etta, G.; Sangiorgio, E.; Alipour, S.; Bonetti, A.; Alvisi, L.; Scala, A.; Baronchelli, A.; Cinelli, M.; and Quattrociocchi, W. 2024. Persistent interaction patterns across social media platforms and over time. *Nature*, 628: 582 – 589.

Balog, K.; Palowitch, J.; Ikica, B.; Radlinski, F.; Alvari, H.; and Manshadi, M. 2024. Towards Realistic Synthetic User-Generated Content: A Scaffolding Approach to Generating Online Discussions. arXiv:2408.08379.

Banker, R. D.; and Datar, S. M. 1989. Software Complexity and Maintainability. In *Proceedings of the 10th International Conference on Information Systems (ICIS)*. Association for Information Systems, Boston, MA: AIS Electronic Library (AISeL).

Beck, K. 2000. *Extreme programming explained: embrace change*. addison-wesley professional.

Bisbee, J.; Clinton, J. D.; Dorff, C.; Kenkel, B.; and Larson, J. M. 2024. Synthetic Replacements for Human Survey Data? The Perils of Large Language Models. *Political Analysis*, 32(4): 401–416.

Burton, J. W.; Lopez-Lopez, E.; Hechtlinger, S.; et al. 2024. How Large Language Models Can Reshape Collective Intelligence. *Nature Human Behaviour*, 8: 1643–1655.

Chang, J. P.; and Danescu, C. 2019. Trouble on the Horizon: Forecasting the Derailment of Online Conversations as they Develop. In Inui, K.; Jiang, J.; Ng, V.; and Wan, X., eds., *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 4743–4754. Hong Kong, China: Association for Computational Linguistics.

Cho, H.; Liu, S.; Shi, T.; Jain, D.; Rizk, B.; Huang, Y.; Lu, Z.; Wen, N.; Gratch, J.; Ferrara, E.; and May, J. 2024. Can Language Model Moderators Improve the Health of Online Discourse? In Duh, K.; Gomez, H.; and Bethard, S., eds., *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 7478–7496. Mexico City, Mexico.

Clemons, E. K.; Schreieck, M.; and Waran, R. V. 2025. Managing disinformation on social media platforms. *Electronic Markets*, 35(52).

Cresci, S.; Trujillo, A.; and Fagni, T. 2022. Personalized Interventions for Online Moderation. In *Proceedings of the 33rd ACM Conference on Hypertext and Social Media*, HT '22, 248–251. New York, NY, USA: Association for Computing Machinery. ISBN 9781450392334.

De Kock, C.; Stafford, T.; and Vlachos, A. 2022. How to disagree well: Investigating the dispute tactics used on Wikipedia. In Goldberg, Y.; Kozareva, Z.; and Zhang, Y., eds., *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 3824–3837. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics.

Denniss, E.; and Lindberg, R. 2025. Social media and the spread of misinformation: infectious and a threat to public health. *Health Promotion International*, 40.

eRulemaking Initiative, C. 2017. CeRI (Cornell e-Rulemaking) Moderator Protocol. Cornell e-Rulemaking Initiative Publications, 21.

Esau, K.; Friess, D.; and Eilders, C. 2017. Design matters! An empirical analysis of online deliberation on different news platforms. *Policy & Internet*, 9(3): 321–342.

Falk, N.; Jundi, I.; Vecchi, E. M.; and Lapesa, G. 2021. Predicting Moderation of Deliberative Arguments: Is Argument Quality the Key? In Al-Khatib, K.; Hou, Y.; and Stede, M., eds., *Proceedings of the 8th Workshop on Argument Mining*, 133–141. Punta Cana, Dominican Republic: Association for Computational Linguistics.

Falk, N.; Vecchi, E.; Jundi, I.; and Lapesa, G. 2024. Moderation in the Wild: Investigating User-Driven Moderation in Online Discussions. In Graham, Y.; and Purver, M., eds., *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, 992–1013. St. Julian's, Malta: Association for Computational Linguistics.

Gligori'c, K.; Zrnic, T.; Lee, C.; Candes, E. J.; and Jurafsky, D. 2024. Can Unconfident LLM Annotations Be Used for Confident Conclusions? *ArXiv*, abs/2408.15204.

Grossmann, I.; Feinberg, M.; Parker, D.; Christakis, N.; Tetlock, P.; and Cunningham, W. 2023. AI and the transformation of social science research. *Science (New York, N.Y.)*, 380: 1108–1109.

Habernal, I.; and Gurevych, I. 2016. Which argument is more convincing? Analyzing and predicting convincingness of Web arguments using bidirectional LSTM. In Erk, K.; and Smith, N. A., eds., *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1589–1599. Berlin, Germany: Association for Computational Linguistics.

Hewitt, L.; Ashokkumar, A.; Ghezae, I.; and Willer, R. 2024. Predicting Results of Social Science Experiments Using Large Language Models. Equal contribution, order randomized.

Horta Ribeiro, M.; Cheng, J.; and West, R. 2023. Automated Content Moderation Increases Adherence to Community Guidelines. In *Proceedings of the ACM Web Conference 2023*, WWW '23, 2666–2676. New York, NY, USA: Association for Computing Machinery. ISBN 9781450394161.

Huang, S.; Siddarth, D.; Lovitt, L.; Liao, T. I.; Durmus, E.; Tamkin, A.; and Ganguli, D. 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, 1395–1417. New York, NY, USA: Association for Computing Machinery. ISBN 9798400704505.

Jaidka, K. 2022. Talking politics: Building data-driven lexica to measure political discussion quality. *Computational Communication Research*, 4(2): 486–527.

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

Kang, H.; and Qian, T. 2024. Implanting LLM's Knowledge via Reading Comprehension Tree for Toxicity Detection. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Findings of the Association for Computational Linguistics ACL 2024*, 947–962. Bangkok, Thailand and virtual meeting: Association for Computational Linguistics.

Kim, S.; Eun, J.; Seering, J.; and Lee, J. 2021. Moderator Chatbot for Deliberative Discussion: Effects of Discussion Structure and Discussant Facilitation. *Proc. ACM Hum.-Comput. Interact.*, 5(CSCW1).

Kolluri, A.; Murthy, D.; and Vinton, K. 2025. Quantifying the spread of racist content on fringe social media: A case study of Parler. *Big Data & Society*, 12(2).

Korre, K.; Tsirmpas, D.; Gkoumas, N.; Cabalé, E.; Kontarinis, D.; Myrtzani, D.; Evgeniou, T.; Androutsopoulos, I.; and Pavlopoulos, J. 2025. Evaluation and Facilitation of Online Discussions in the LLM Era: A Survey. ACL ARR 2025 February Submission.

Kumar, D.; AbuHashem, Y. A.; and Durumeric, Z. 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.

Leng, Y.; and Yuan, Y. 2024. Do LLM Agents Exhibit Social Behavior? arXiv:2312.15198.

Li, A.; Zhou, Y.; Raghuram, V. C.; Goldstein, T.; and Goldblum, M. 2025. Commercial LLM Agents Are Already Vulnerable to Simple Yet Dangerous Attacks. arXiv:2502.08586.

Lin, C.-Y. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, 74–81. Barcelona, Spain: Association for Computational Linguistics.

Lukito, J.; Chen, B.; Masullo, G. M.; and Stroud, N. J. 2024. Comparing a BERT Classifier and a GPT classifier for Detecting Connective Language Across Multiple Social Media. In Al-Onaizan, Y.; Bansal, M.; and Chen, Y.-N., eds., *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 19140–19153. Miami, Florida, USA: Association for Computational Linguistics.

Majumdar, D.; S, A.; Boyina, P.; Rayidi, S. S. P.; Sai, Y. R.; and Gangashetty, S. V. 2024. Beyond Text: Nefarious Actors Harnessing LLMs for Strategic Advantage. In *2024 International Conference on Intelligent Systems for Cybersecurity (ISCS)*, 1–7.

Marulli, F.; Paganini, P.; and Lancellotti, F. 2024. The Three Sides of the Moon LLMs in Cybersecurity: Guardians, Enablers and Targets. *Procedia Computer Science*, 246: 5340–5348. 28th International Conference on Knowledge Based and Intelligent information and Engineering Systems (KES 2024).

Marzo, G. D.; Pietronero, L.; and Garcia, D. 2023. Emergence of Scale-Free Networks in Social Interactions among Large Language Models. arXiv:2312.06619.

Mou, X.; Wei, Z.; and Huang, X. 2024. Unveiling the Truth and Facilitating Change: Towards Agent-based Large-scale Social Movement Simulation. arXiv:2402.16333.

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

Neumann, T.; De-Arteaga, M.; and Fazelpour, S. 2025. Should you use LLMs to simulate opinions? Quality checks for early-stage deliberation. arXiv:2504.08954.

Nik Ahmad, N. A. 2010. CETLs : Supporting Collaborative Activities Among Students and Teachers Through the Use of Think- Pair-Share Techniques. *International Journal of Computer Science Issues*, 7.

Ogheneovo, E. E.; et al. 2014. On the relationship between software complexity and maintenance costs. *Journal of Computer and Communications*, 2(14): 1.

OpenAI. 2024. GPT-4 Technical Report. arXiv:2303.08774.

Park, J. S.; O'Brien, J. C.; Cai, C. J.; Morris, M. R.; Liang, P.; and Bernstein, M. S. 2023a. Generative Agents: Interactive Simulacra of Human Behavior. arXiv:2304.03442.

Park, J. S.; O'Brien, J. C.; Cai, C. J.; Morris, M. R.; Liang, P.; and Bernstein, M. S. 2023b. Generative Agents: Interactive Simulacra of Human Behavior. *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*.

Park, J. S.; Popowski, L.; Cai, C.; Morris, M. R.; Liang, P.; and Bernstein, M. S. 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. ISBN 9781450393201.

Park, J. S.; Zou, C. Q.; Shaw, A.; Hill, B. M.; Cai, C.; Morris, M. R.; Willer, R.; Liang, P.; and Bernstein, M. S. 2024. Generative Agent Simulations of 1,000 People. arXiv:2411.10109.

Pavlopoulos, J.; and Likas, A. 2024. Polarized Opinion Detection Improves the Detection of Toxic Language. In Graham, Y.; and Purver, M., eds., *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1946–1958. St. Julian's, Malta: Association for Computational Linguistics.

Pavlopoulos, J.; Sorensen, J.; Dixon, L.; Thain, N.; and Androutsopoulos, I. 2020. Toxicity Detection: Does Context Really Matter? In Jurafsky, D.; Chai, J.; Schluter, N.; and Tetreault, J., eds., *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 4296–4305. Online: Association for Computational Linguistics.

Persing, I.; and Ng, V. 2015. Modeling Argument Strength in Student Essays. In Zong, C.; and Strube, M., eds., *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)*, 543–552. Beijing, China: Association for Computational Linguistics.

Pranesh, S.; and Gupta, S. 2024. The impact of social media on polarization in the society. arXiv:2408.12877.

Rooderkerk, R. P.; and Pauwels, K. H. 2016. No Comment?! The Drivers of Reactions to Online Posts in Professional Groups. *Journal of Interactive Marketing*, 35(1): 1–15.

Rosenberg, M. B.; and Chopra, D. 2015. *Nonviolent communication: A language of life: Life-changing tools for healthy relationships*. PuddleDancer Press.

Rossetti, G.; Stella, M.; Cazabet, R.; Abramski, K.; Cau, E.; Citraro, S.; Failla, A.; Improta, R.; Morini, V.; and Pansanella, V. 2024. Y Social: an LLM-powered Social Media Digital Twin. arXiv:2408.00818.

Rossi, L.; Harrison, K.; and Shklovski, I. 2024. The Problems of LLM-generated Data in Social Science Research. *Sociologica*, 18(2): 145–168.

Schaffner, B.; Bhagoji, A. N.; Cheng, S.; Mei, J.; Shen, J. L.; Wang, G.; Chetty, M.; Feamster, N.; Lakier, G.; and Tan, C. 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. ISBN 9798400703300.

Schluger, C.; Chang, J.; Danescu-Niculescu-Mizil, C.; and Levy, K. 2022. Proactive Moderation of Online Discussions: Existing Practices and the Potential for Algorithmic Support. *Proc. ACM Hum.-Comput. Interact.*, 6(CSCW2).

Schroeder, H.; Roy, D.; and Kabbara, J. 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*, 13985–14001. Bangkok, Thailand.

Seering, J. 2020. Reconsidering Self-Moderation: the Role of Research in Supporting Community-Based Models for Online Content Moderation. *Proc. ACM Hum.-Comput. Interact.*, 4(CSCW2).

Small, C. T.; Vendrov, I.; Durmus, E.; Homaei, H.; Barry, E.; Cornebise, J.; Suzman, T.; Ganguli, D.; and Megill, C. 2023. Opportunities and Risks of LLMs for Scalable Deliberation with Polis. *ArXiv*, abs/2306.11932.

Taubenfeld, A.; Dover, Y.; Reichart, R.; and Goldstein, A. 2024. Systematic Biases in LLM Simulations of Debates. *ArXiv*, abs/2402.04049.

Thomas, D. 2025. *simplicity: sustainable, humane, and effective software development*. The Pragmatic Programmers LLC.

Tsai, L. L.; Pentland, A.; Braley, A.; Chen, N.; Enríquez, J. R.; and Reuel, A. 2024. Generative AI for Pro-Democracy Platforms. *An MIT Exploration of Generative AI*. Https://mit-genai.pubpub.org/pub/mn45hexw.

Tsai, V.; Qian, C.; and comm. contributors, D. L. 2025. Deliberate Lab: Open-Source Platform for LLM-Powered Social Science.

Törnberg, P.; Valeeva, D.; Uitermark, J.; and Bail, C. 2023. Simulating Social Media Using Large Language Models to Evaluate Alternative News Feed Algorithms. arXiv:2310.05984.

Ulmer, D.; Mansimov, E.; Lin, K.; Sun, J.; Gao, X.; and Zhang, Y. 2024. Bootstrapping LLM-based Task-Oriented Dialogue Agents via Self-Talk. *ArXiv*, abs/2401.05033.

Vezhnevets, A. S.; Agapiou, J. P.; Aharon, A.; Ziv, R.; Matyas, J.; Du'enez-Guzm'an, E. A.; Cunningham, W. A.; Osindero, S.; Karmon, D.; and Leibo, J. Z. 2023. Generative agent-based modeling with actions grounded in physical, social, or digital space using Concordia. *ArXiv*, abs/2312.03664.

Wachsmuth, H.; Naderi, N.; Hou, Y.; Bilu, Y.; Prabhakaran, V.; Thijm, T. A.; Hirst, G.; and Stein, B. 2017. Computational Argumentation Quality Assessment in Natural Language. In Lapata, M.; Blunsom, P.; and Koller, A., eds., *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, 176–187. Valencia, Spain: Association for Computational Linguistics.

Wang, H.; Fu, W.; Tang, Y.; Chen, Z.; Huang, Y.; Piao, J.; Gao, C.; Xu, F.; Jiang, T.; and Li, Y. 2025. A Survey on Responsible LLMs: Inherent Risk, Malicious Use, and Mitigation Strategy. arXiv:2501.09431.

Wang, Y.-S.; and Chang, Y. T. 2022. Toxicity Detection with Generative Prompt-based Inference. *ArXiv*, abs/2205.12390.

White, K.; Hunter, N.; and Greaves, K. 2024. *facilitating deliberation - a practical guide*. Mosaic Lab.

Xia, Y.; Zhu, H.; Lu, T.; Zhang, P.; and Gu, N. 2020. Exploring Antecedents and Consequences of Toxicity in Online Discussions: A Case Study on Reddit. *Proc. ACM Hum.-Comput. Interact.*, 4(CSCW2).

Yu, Y.; Yao, Z.; Li, H.; Deng, Z.; Cao, Y.; Chen, Z.; Suchow, J. W.; Liu, R.; Cui, Z.; Xu, Z.; Zhang, D.; Subbalakshmi, K.; Xiong, G.; He, Y.; Huang, J.; Li, D.; and Xie, Q. 2024. FinCon: A Synthesized LLM Multi-Agent System with Conceptual Verbal Reinforcement for Enhanced Financial Decision Making. arXiv:2407.06567.

Ziegele, M.; Weber, M.; Quiring, O.; and and, T. B. 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.

Reproducibility Checklist

Instructions for Authors:

This document outlines key aspects for assessing reproducibility. Please provide your input by editing this `.tex` file directly.

For each question (that applies), replace the "Type your response here" text with your answer.

Example: If a question appears as

```
\question{Proofs of all novel claims
are included} {(yes/partial/no)}
Type your response here
```

you would change it to:

```
\question{Proofs of all novel claims
are included} {(yes/partial/no)}
yes
```

Please make sure to:

- Replace ONLY the "Type your response here" text and nothing else.
- Use one of the options listed for that question (e.g., **yes**, **no**, **partial**, or **NA**).
- **Not** modify any other part of the `\question` command or any other lines in this document.

You can `\input` this .tex file right before `\end{document}` of your main file or compile it as a stand-alone document. Check the instructions on your conference's website to see if you will be asked to provide this checklist with your paper or separately.

---

1. General Paper Structure

1.1. Includes a conceptual outline and/or pseudocode description of AI methods introduced (yes/partial/no/NA) yes

1.2. Clearly delineates statements that are opinions, hypothesis, and speculation from objective facts and results (yes/no) yes

1.3. Provides well-marked pedagogical references for less-familiar readers to gain background necessary to replicate the paper (yes/no) yes?

2. Theoretical Contributions

2.1. Does this paper make theoretical contributions? (yes/no) no

If yes, please address the following points:

2.2. All assumptions and restrictions are stated clearly and formally (yes/partial/no)

2.3. All novel claims are stated formally (e.g., in theorem statements) (yes/partial/no)

2.4. Proofs of all novel claims are included (yes/partial/no)

2.5. Proof sketches or intuitions are given for complex and/or novel results (yes/partial/no)

2.6. Appropriate citations to theoretical tools used are given (yes/partial/no)

2.7. All theoretical claims are demonstrated empirically to hold (yes/partial/no/NA)

2.8. All experimental code used to eliminate or disprove claims is included (yes/no/NA)

3. Dataset Usage

3.1. Does this paper rely on one or more datasets? (yes/no) yes

If yes, please address the following points:

3.2. A motivation is given for why the experiments are conducted on the selected datasets (yes/partial/no/NA) yes

3.3. All novel datasets introduced in this paper are included in a data appendix (yes/partial/no/NA) yes

3.4. All novel datasets introduced in this paper will be made publicly available upon publication of the paper with a license that allows free usage for research purposes (yes/partial/no/NA) yes

3.5. All datasets drawn from the existing literature (potentially including authors' own previously published work) are accompanied by appropriate citations (yes/no/NA) yes

3.6. All datasets drawn from the existing literature (potentially including authors' own previously published work) are publicly available (yes/partial/no/NA) yes

3.7. All datasets that are not publicly available are described in detail, with explanation why publicly available alternatives are not scientifically satisfying (yes/partial/no/NA) NA

4. Computational Experiments

4.1. Does this paper include computational experiments? (yes/no) yes

If yes, please address the following points:

4.2. This paper states the number and range of values tried per (hyper-) parameter during development of the paper, along with the criterion used for selecting the final parameter setting (yes/partial/no/NA) NA

4.3. Any code required for pre-processing data is included in the appendix (yes/partial/no) yes

4.4. All source code required for conducting and analyzing the experiments is included in a code appendix (yes/partial/no) yes

4.5. All source code required for conducting and analyzing the experiments will be made publicly available upon publication of the paper with a license that allows free usage for research purposes (yes/partial/no) yes

4.6. All source code implementing new methods have comments detailing the implementation, with references to the paper where each step comes from (yes/partial/no) yes

4.7. If an algorithm depends on randomness, then the method used for setting seeds is described in a way sufficient to allow replication of results (yes/partial/no/NA) yes

4.8. This paper specifies the computing infrastructure used for running experiments (hardware and software), including GPU/CPU models; amount of memory; operating system; names and versions of relevant software libraries and frameworks (yes/partial/no) yes

4.9. This paper formally describes evaluation metrics used and explains the motivation for choosing these metrics (yes/partial/no) yes

4.10. This paper states the number of algorithm runs used to compute each reported result (yes/no) yes

4.11. Analysis of experiments goes beyond single-dimensional summaries of performance (e.g., average; median) to include measures of variation, confidence, or other distributional information (yes/no) no

4.12. The significance of any improvement or decrease in performance is judged using appropriate statistical tests (e.g., Wilcoxon signed-rank) (yes/partial/no) yes

4.13. This paper lists all final (hyper-)parameters used for each model/algorithm in the paper's experiments (yes/partial/no/NA) NA

Figure 7: Difference in average Arg. Quality 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 ($\cdot p < 0.1$, $^{*}\ p < 0.05$, $^{**}\ p < 0.01$, $^{***}\ p < 0.001$).

Technical Appendix

Generalizing Synthetic Annotation

While toxicity is a reliable and important metric, we can also investigate other discussion quality dimensions, such as Arg. Quality. Arg. Quality is an important metric, frequently studied in the field of online facilitation (Argyle et al. 2023; Schroeder, Roy, and Kabbara 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 Arg. Quality 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 Arg. Quality(§4).

Most findings w.r.t. toxicity are mirrored for Arg. Quality. Fig. 7 demonstrates that the presence of an LLM facilitator qualitatively improves the Arg. Quality of synthetic discussions, although to a lesser extent when compared with toxicity (c.f. Fig. 3). Similarly, there is no qualitative, observed improvement when advanced facilitation strategies are used (Fig. 7). LLM users also show worse Arg. Quality in the presence of trolls, when we use our specialized instruction prompt. Contrary to toxicity, the presence of LLM facilitators does not seem to improve Arg. Quality over time, as demonstrated in Table 3.

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^{***}

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

Table 3: OLS regression coefficients for Arg.Q. ($Adj.R^2 = 0.016$). *"Time"* denotes dialogue turn, reference factor is *No Moderator*.

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

Our analysis reveals a positive correlation between toxicity and annotator polarization: As demonstrated by Fig. 10, while there is general agreement on non-toxic comments, annotators struggle to reach consensus as toxicity becomes non-trivial (*toxicity* $\in [2, 5]$) with a statistically significant difference (Student's t-test $p < .000$). This phenomenon does not manifest in the Arg. Quality scores.

To mitigate the instability inherent in LLM outputs—even when given identical inputs—the use of multiple annotator-agents is essential for obtaining reliable annotations. To demonstrate this necessity, we run an experiment where we use ten annotator-agents on a subset of comments with the same annotator model and instruction prompt, but no SDBs. As illustrated in Fig. 9, even under conditions which guaranteed identical inputs, there exists some polarization, with some comments even showing maximum polarization. Running the same experiment with different SDBs yields identical results, indicating that the observed polarization is primarily due to unstable model outputs. Thus, we confirm the results of previous studies on LLM instability (Rossi, Harrison, and Shklovski 2024; Atil et al. 2025), while also bypassing this limitation in our own results.

Investigating Confounding Variables

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

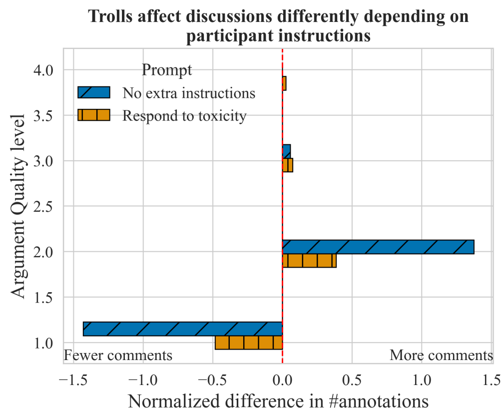


Figure 8: Relative differences in number of annotations per Arg. Quality of synthetic discussions, when comments by troll users are excluded. We compare between our specialized and a basic instruction prompt.

comments ($p = 0.775$). This may partially explain the divergent diversity scores of different models in §5.2 (Fig. 12a) — although it does not ultimately change the conclusions presented. Furthermore, Fig. 11 shows that toxicity and Arg. Quality are on average not qualitatively dependent on the model used.

Prompts Used

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.

User Roles

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

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

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

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:

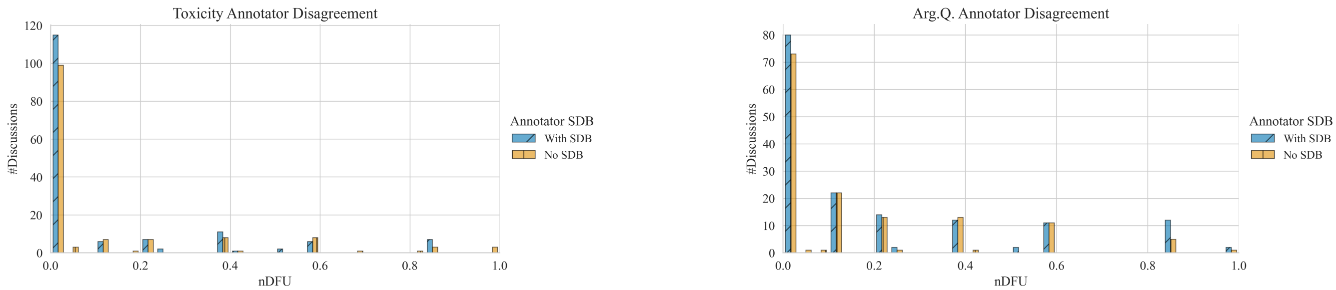


Figure 9: Distribution plot of inter-annotator polarization (nDFU) for each comment in all synthetic discussions following the ”No Instructions” strategy and using the Qwen 2.5 model. The blue (left-most) bars represent the disagreement between 10 identical annotator-agents, while the orange (right-most)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

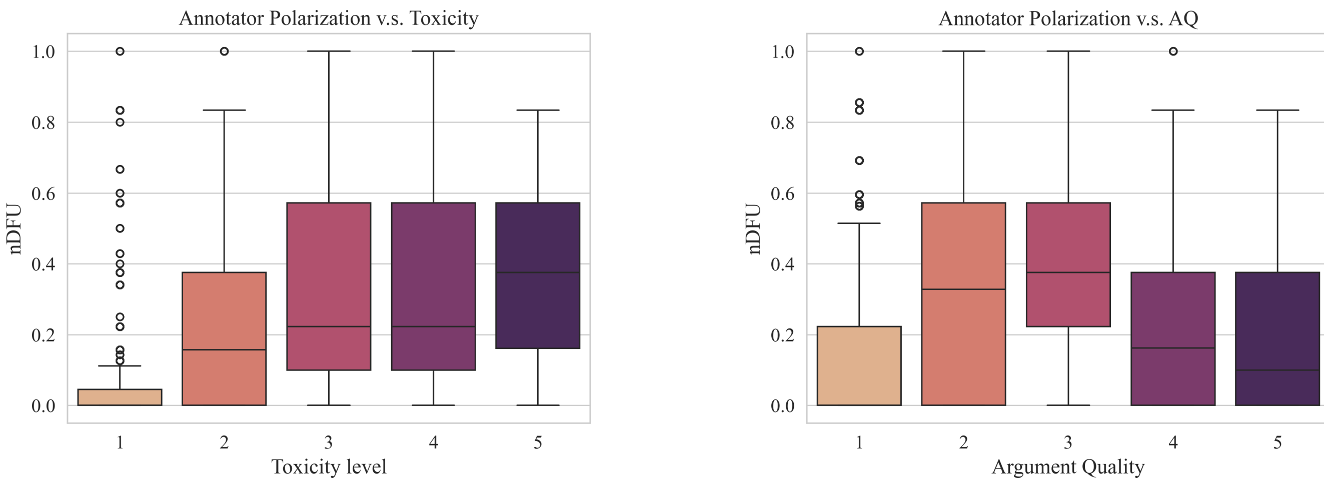


Figure 10: Inter-annotator polarization (nDFU) of each synthetic comment for all synthetic discussions, by annotation level. The left graph shows the relationship between $nDFU_{toxicity}$ and toxicity, while the right graph shows the relationship between $nDFU_{arg\_quality}$ and Arg. Quality.

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

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

- **Regulation Room**: Oversee the conversation and ensure the rules below are enforced. Follow the following guidelines:
  - **Encourage Informed Commenting**: Guide users to

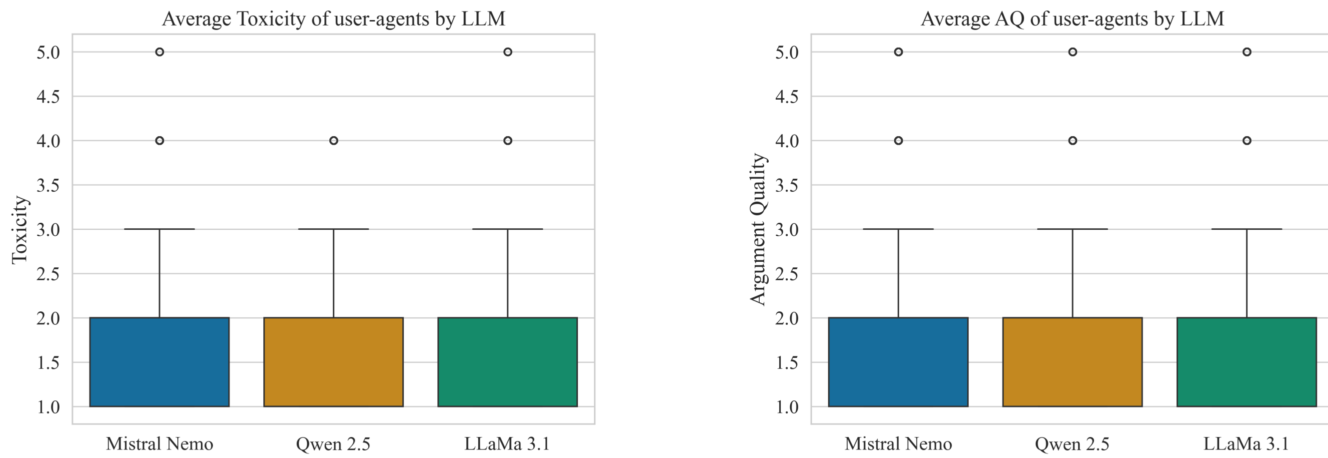
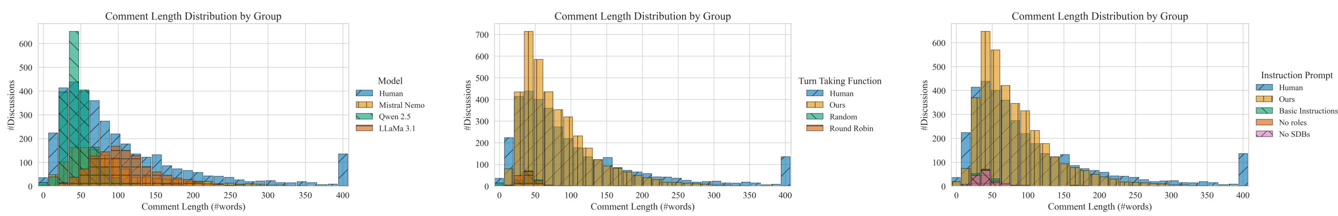


Figure 11: Boxplots for average Toxicity (left) and Arg. Quality (right) per LLM (§4.4).



(a) Model

(b) Turn-taking function t

(c) Prompting function ϕ

Figure 12: Comment length for each discussion by LLM (§4.4), turn-taking function t, and prompt used (§3.2). For ease of comparison, comments above 400 words are marked at the end of the x-axis.

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.