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 datasetcontaining 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[1] that efficiently implements our methodology at great scale.

1 Introduction

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

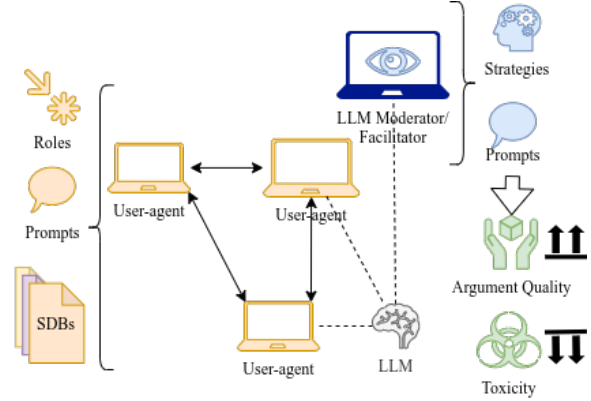[1]anonymous.4open.science/r/framework-F8E6

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

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

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

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

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

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

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

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 classification models traditionally used in online platforms, LLMs can actively facilitate discussions (Korre et al., 2025). They can warn users for rule violations (Kumar et al., 2024), monitor engagement (Schroeder et al., 2024), aggregate diverse opinions (Small et al., 2023), and provide translations and writing tips, which is especially useful for marginalized groups (Tsai et al., 2024). These capabilities suggest that LLMs may be able to assist or even replace human facilitators in many tasks (Small et al., 2023; Seering, 2020).

Moderator chatbots have shown promise; Kim et al. (2021) demonstrated that simple rule-based

models can enhance discussions, although their approach was largely confined to organizing the discussion based on the "think-pair-share" framework (Nik Ahmad, 2010; Navajas et al., 2018), and balancing user activity. Cho et al. (2024) use LLM facilitators in human discussions, with facilitation strategies based on Cognitive Behavioral Therapy and the work of Rosenberg and Chopra (2015). They show that LLM facilitators can provide "specific and fair feedback" to users, although they struggle to make users more respectful and cooperative. In contrast to both works, our work uses exclusively LLM participants and LLM facilitators, and tests the latter in an explicitly toxic and challenging environment.

2.3 Discussion Quality

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

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

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

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

3.2 Turn Taking

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

3.3 Instruction Prompting

We use a standard instruction prompt for the non-facilitator participants, which instructs them to respond to repeatedly toxic comments. This was a necessary measure to bypass the extreme agreeableness of LLMs (Park et al., 2023).

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., help-

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. Comments are clipped due to length. @CynicalInvestor88 is also a part of the discussion; not a hallucination.

ing the community or disrupting discussions). Each role was mapped to specific instructions. We create three roles for users: neutral users, trolls, and community veterans.

3.4 LLM Personas

SocioDemographic Backgrounds (SDBs) have proven promising in generating varied responses from LLMs, and alleviating the Western bias exhibited by them (Burton et al., 2024). We generate 30 LLM user personas with unique SDBs by prompting a GPT-4 model (OpenAI et al., 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 Facilitation Strategies

We test four different facilitation strategies, 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. Example: "You are a moderator, keep the discussion civil".
3. **Rules Only**: A *real-life* strategy where the prompt is adapted from LLM alignment guidelines (Huang et al., 2024). These guidelines were selected to be as unanimously agreed upon across various human groups. They thus provide a set of rules to uphold, without specifying *how* to uphold them (e.g, "Be fair and impartial, assist users, don't spread misinformation").

4. **Regulation Room**: A *real-life* strategy based on guidelines given to human facilitators of the "Regulation Room" platform (eRulemaking Initiative, 2017). The instructions are typical of online moderation. Example: "Stick to a maximum of two questions, use simple and clear language, deal with off-topic comments".

5. **Constructive Communications**: A *real-life* strategy based on the human facilitation guidelines used by the MIT Center for Constructive Communications (White et al., 2024). It approaches facilitation from a more personalized and indirect angle, forbidding facilitators from directly providing opinions or directions. Example: "Do not make decisions, be a guide, provide explanations".

6. **Moderation Game**: Our proposed *experimental* strategy, inspired by Abdelnabi et al. (2024) (see §2.1). Instructions are formulated as a game, where the facilitator LLM tries to maximize their scores by arriving at specific outcomes. No actual score is being kept; they exist to act as indications for how desirable an outcome is. The other participants are not provided with scores, nor are they aware of the game rules. Example: "User is toxic: -5 points, User corrects behavior: $+10$ points".

4.2 Synthetic Discussion Generation

An overview of how the experiments are generated can be found in Algorithm 1. We provide our framework with a set of starting opinions ("seed opinions") and SDBs. We then run $N_d = 8$ discussions for each pair of facilitation strategies S and model. Synthetic generation is then handled as described in §3.

4

Algorithm 1 Synthetic discussion setup generation

> **Input:**
> - User SDBs $\Theta = \{\theta_1, \dots, \theta_{30}\}$
> - Moderator SDB = θ_{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: $D = D \cup d$
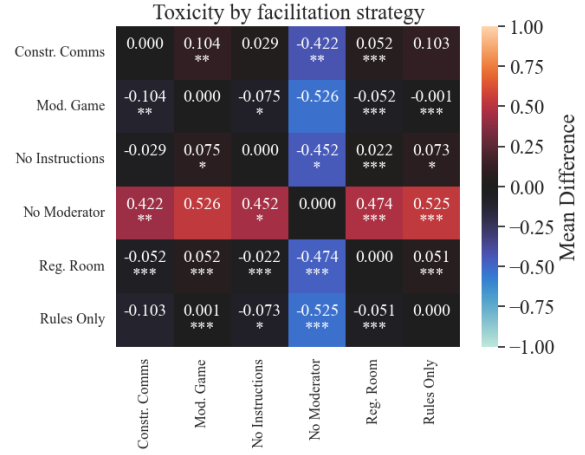> 11: **return** D

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

4.3 Evaluation

In our study, we use *toxicity* as a proxy for discussion quality, since it can inhibit online and deliberative discussions (De Kock et al., 2022; Xia et al., 2020)[2]. We use ten LLM annotator-agents controlled by a model already used in prior work (LLaMa3.1 70B) (Kang and Qian, 2024) (§4.3), as LLMs are reliable for toxicity detection (Kang and Qian, 2024; Wang and Chang, 2022; Anjum and Katarya, 2024).

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

We note again that these metrics are better 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: LLaMa 3.2 (70B), Qwen2.5 (33B), Mistral Nemo (12B), quantized to 4 bits. All the experiments were collectively completed within four weeks of computational time, using two Quadro RTX 6000 GPUs. The process of generating discussion setups is detailed in §4.2. The execution script is available in the project's repository.[3]

5 Results

5.1 Main findings

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

Finding 2: More elaborate facilitation strategies fail to decrease toxicity. More elaborate facilitation strategies, such as *Regulation Room*, *Constructive Communications*, and our proposed *Moderation Game*, lead to a statistically significant reduction in comment toxicity over time compared

---

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

[3]anonymous.4open.science/r/experiments-B27D

[4]The 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***

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

Table 1: OLS regression coefficients for toxicity ($Adj.R^2 = 0.054$). Average toxicity without facilitators is 2.164.

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

to *unmoderated* discussions (Table 1). However, their additional impact beyond that of the simpler *No Instructions* strategy is marginal and sometimes not statistically significant (Fig. 3), suggesting that out-of-the-box LLMs may struggle to effectively leverage advanced instructions—echoing prior findings on the limitations of LLM facilitators (Cho et al., 2024).

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

Additionally, a qualitative look through the dataset reveals that LLM user-agents exhibit 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). This is likely another artifact caused by alignment procedures, making LLMs too agreeable (Park et al., 2023).
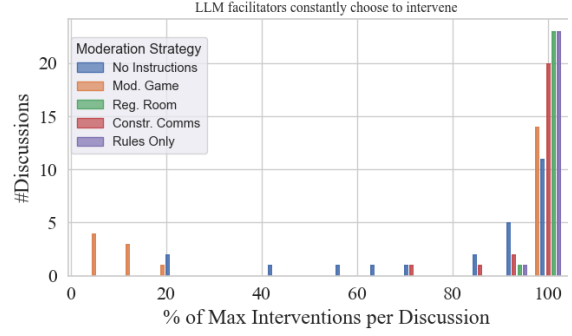
5.2 Ablation Study

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 ones from the CeRI "Regulation Room" dataset[5], which includes moderated online deliberative discussions for ten diverse topics.

5.2.1 Effects of LLMs

Larger models do not lead to more high-quality discussions. As shown in Fig. 5a, Qwen demonstrated the highest diversity among the evaluated models, indicating limited participant interaction (§2.3), followed by Mistral Nemo and LLaMa. However, none of the models closely matched the diversity observed in human discussions.

5.2.2 Effects of Turn-Taking Functions

Our proposed turn-taking function substantially improves the quality of synthetic data. We compare our turn-taking function (§3.2) to two baselines: Round Robin (participants speaking one after the other, then repeating) and Random Selection (uniformly sampling another participant each turn). Fig. 5b demonstrates 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).

5.2.3 Effects of User Prompting

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

---

[5] http://archive.regulationroom.org. Disclaimer: Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the CeRI.
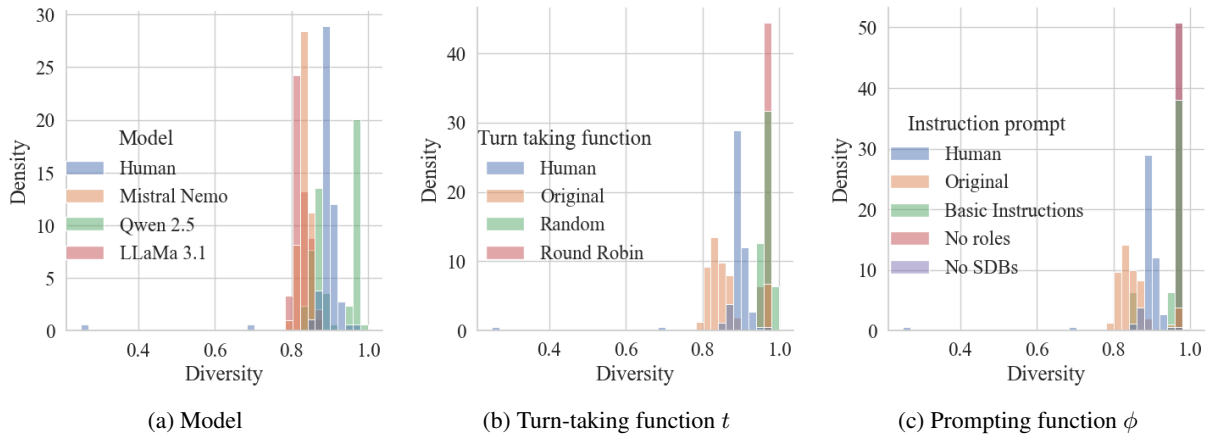
(a) Model   (b) Turn-taking function t (c) Prompting function ϕ

Figure 5: Diversity (§2.3) distribution for each discussion by LLM (§4.4), turn-taking function t (§3.2), and prompting function ϕ used (§3.3).
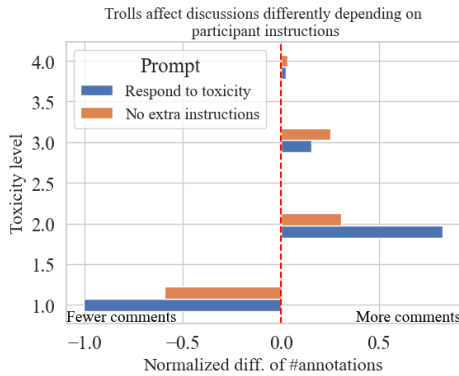

Figure 6: Non-troll toxicity levels in discussions with and without trolls. There is a significant uptick on the number of "somewhat toxic" ($Toxicity = 2$) comments when the participants are primed to respond to toxic comments.

assigned roles, or (3) only a very basic instruction prompt given.

Specialized instruction prompts are essential for eliciting toxic behavior in instruction-tuned LLMs. Our instruction prompt for the participants (§3.3) incentivizes them to react to toxic behavior. Indeed, inserting "troll" participants to discussions, leads to more intense toxicity among *other* participants *only if we instruct participants to react to toxic posts* (Fig. 6).

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

6 Datasets and Software

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

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

We also release a dataset of synthetic discussions annotated by LLMs. It can serve for finetuning facilitator LLMs, as generally showcased by Ulmer et al. (2024). The supplementary ablation dataset, as well as the code for the analysis and the graphs present in this paper, can be found in the project repository[7]. 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 Conclusions

We proposed a simple and generalizable methodology that enables researchers to quickly and inexpensively conduct pilot facilitation experiments using exclusively LLMs and validated it through an ablation study. We found that (1) LLM facilitators significantly improve the quality of synthetic discussions; (2) prompting these facilitators with strategies based on Social Science research

---

[6]anonymous.4open.science/r/framework-F8E6
[7]anonymous.4open.science/r/experiments-B27D

does not markedly improve their performance. We also identified a consistent problem with LLMs not keeping silence when appropriate. Finally, we created an open-source Python Framework that applies this methodology to hundreds of experiments, which we used to create and publish a large-scale synthetic dataset.

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.

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.

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

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

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

Krisztian Balog, John Palowitch, Barbara Ikica, Filip Radlinski, Hamidreza Alvari, and Mehdi Manshadi. 2024. Towards realistic synthetic user-generated content: A scaffolding approach to generating online discussions. *Preprint*, arXiv:2408.08379.

J. W. Burton, E. Lopez-Lopez, S. Hechtlinger, and 1 others. 2024. How large language models can reshape collective intelligence. *Nature Human Behaviour*, 8:1643–1655.

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

Erik K. Clemons, Matthias Schreieck, and Rishi V. Waran. 2025. Managing disinformation on social media platforms. *Electronic Markets*, 35(52).

Stefano Cresci, Amaury Trujillo, and Tiziano Fagni. 2022. Personalized interventions for online moderation. In *Proceedings of the 33rd ACM Conference on Hypertext and Social Media*, HT '22, page 248–251, New York, NY, USA. Association for Computing Machinery.

Christine De Kock, Tom Stafford, and Andreas Vlachos. 2022. How to disagree well: Investigating the dispute tactics used on Wikipedia. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3824–3837, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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

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

Katharina Esau, Dennis Friess, and Christiane Eilders. 2017. Design matters! an empirical analysis of online deliberation on different news platforms. *Policy & Internet*, 9(3):321–342.

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

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

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

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

8

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

Akaash Kolluri, Dhiraj Murthy, and Kami Vinton. 2025. Quantifying the spread of racist content on fringe social media: A case study of parler. *Big Data & Society*, 12(2).

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

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

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

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

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

Nik Azlina Nik Ahmad. 2010. Cetls : Supporting collaborative activities among students and teachers through the use of think- pair-share techniques. *International Journal of Computer Science Issues*, 7.

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

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. *Preprint*, arXiv:2304.03442.

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

John Pavlopoulos, Jeffrey Sorensen, Lucas Dixon, Nithum Thain, and Ion Androutsopoulos. 2020. Toxicity detection: Does context really matter? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4296–4305, Online. Association for Computational Linguistics.

Samana Pranesh and Sayan Gupta. 2024. The impact of social media on polarization in the society. *Preprint*, arXiv:2408.12877.

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

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

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

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

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

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

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

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

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

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

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

Alexander Sasha Vezhnevets, John P. Agapiou, Avia Aharon, Ron Ziv, Jayd Matyas, Edgar A. Du'enez-Guzm'an, William A. Cunningham, Simon Osindero, Danny Karmon, and Joel Z. Leibo. 2023. Generative agent-based modeling with actions grounded in physical, social, or digital space using concordia. *ArXiv*, abs/2312.03664.

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.