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 a simple, generalizable, LLM-driven methodology to prototype the development of LLM facilitators, and produce high-quality synthetic data without human involvement. We use our methodology to test whether current Social Science strategies for facilitation can improve the performance of LLM facilitators. We find that, while LLM facilitators significantly improve synthetic discussions, there is no evidence that the application of these strategies leads to further improvements in discussion quality. We confirm that each component of our methodology contributes substantially to high quality data via an ablation study. In an effort to aid research in the field of facilitation, we release a large, publicly available dataset containing LLM-generated and LLM-annotated discussions using multiple open-source models. This dataset can be used for LLM facilitator finetuning as well as behavioral analysis of current out-of-the-box LLMs in the task. We also release an open-source python framework that efficiently implements our methodology at great scale.

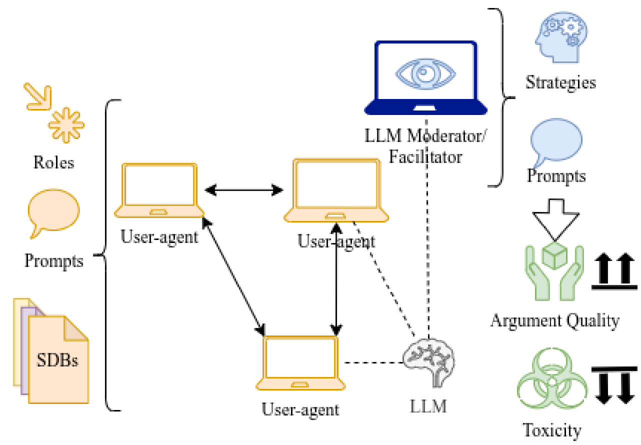


Figure 1: LLM user-agents with distinct SDBs participate in a discussion, while the LLM moderator monitors and attempts to improve the quality of the discussion. We need to design prompts and configurations for both types of LLM agents.

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, 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). 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 (Rossi, Harrison, and Shklovski (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, Harrison, and Shklovski 2024), before testing them with human participants. We propose a simple and generalizable methodology which enables rapid model "debugging" and parameter testing (e.g., discarding sub-optimal prompts for the LLM facilitator) without human involvement (Fig. 1, §3). An ablation study demonstrates that each component of our methodology substantially contributes to generating high-quality data (§5.2).

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

Finally, we release an open-source Python framework that implements our methodology at scale, enabling the research community to rapidly experiment with LLM-based facilitators. We also release a large, publicly available dataset with LLM-generated and annotated synthetic discussions (§6). Our dataset can be used for LLM facilitator finetuning (Ulmer et al. 2024) (given that existing facilitation datasets are few and generally small —Korre et al. (2025)), 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, 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) 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, 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 "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). There are however some metrics 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 metric called *"Diversity"*, which penalizes repeated text sequences between comments using ROUGE-L (Lin 2004) scores. Their approach suffers from the limitations of ROUGE scores (mainly 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. 2023a; 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, and we leave them for future work.

3 Methodology

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

- **Context passing**: How an LLM receives the context of the discussion so far (§3.1).
- **Turn order**: Given that LLMs are trained to be chatbot 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 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, explainable system.

Name	Type
Username	string
Age	integer
	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 project repository for the actual values.

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 participants, while lower values tend to create scenarios with minimal interaction between them. 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. 2023b).

Additionally, following the paradigm presented by Abdelnabi et al. (2024), we assign roles to non-facilitator user-agents, which inform their incentives for participating in the discussion (e.g., helping the community or disrupting discussions). Each role was mapped to specific instructions. 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 (Table 1) 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).

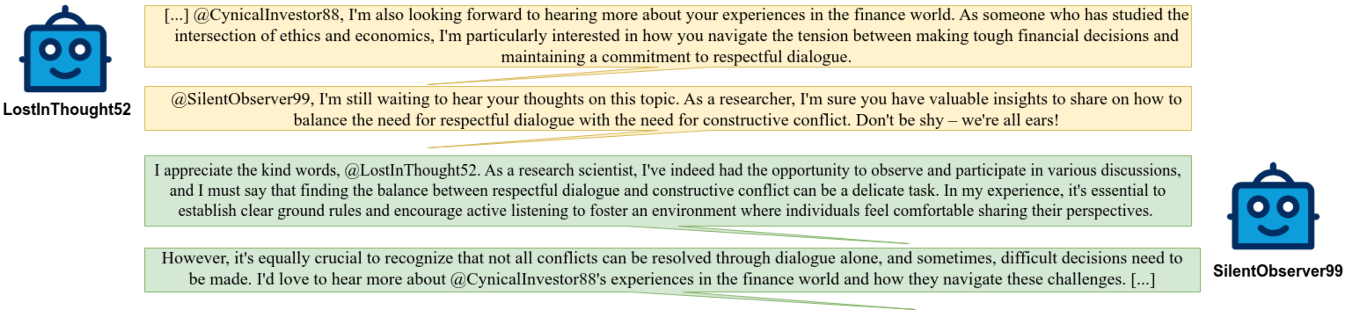


Figure 2: Excerpt from a synthetic discussion. The LLM participants use their sociodemographic prompts to insert personal stories and justify their perspectives in the discussion. Comments are clipped due to length. @CynicalInvestor88 is also a part of the discussion; not a hallucination.

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 = $\{LLaMa, Mistral, Qwen\}$

Output: Set of discussions D

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

4 Experimental Setup

4.1 Synthetic Discussion Generation

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 LLM (§4.4). Synthetic generation is then handled as described in §3.

An overview of how the experiments are generated can be found in Algorithm 1. The $RandomSample$ function returns a number of samples from the uniform distribution. The $ACTORS$ function creates a LLM agent using a model and a (string) prompt.

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. Example: "You are a moderator, keep the discussion civil". This approach is already being used in some platforms (Tsai, Qian, and community contributors 2025).
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"), leaving the LLM completely unconstrained.
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 efficiency (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”.

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, Stafford, and Vlachos 2022; Xia et al. 2020)[1]. 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 (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” (§2.4), we use the “diversity” metric (Ulmer et al. 2024). Low diversity points to pathological problems (e.g., LLMs repeating previous comments). On the other hand, extremely high diversity may point to a lack of interaction between participants; a discussion in which participants engage with each other will feature some lexical overlap (e.g., common terms, paraphrasing points of other participants). We compare the distribution of diversity scores for synthetic discussions with that measured on sampled human discussions. This allows us to estimate the extent to which synthetic discussions approximate real-world content variety and participant interaction.

We note again that these metrics are better 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 §4.1. The execution script is available in the project’s repository.

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.

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 to *unmoderated* discussions (Table 2). However, this is not the case with the *No Instructions* strategy (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).

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

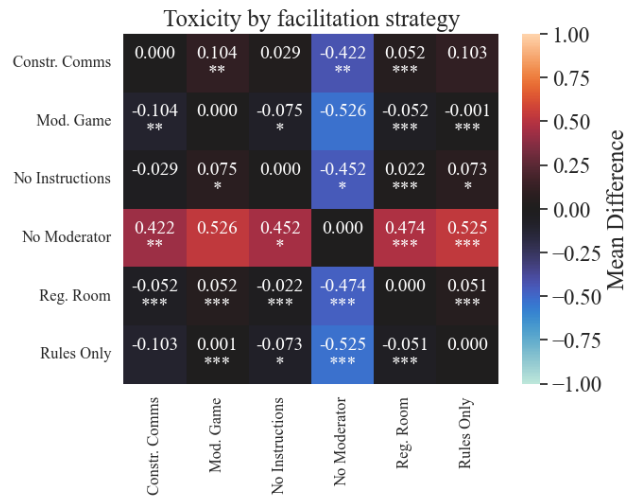


Figure 3: Difference in average toxicity levels for comments following pairs of facilitation strategies. Red cells ($x > 0$) indicate that the strategy on the left performs worse than the one on the bottom, for an average of x points in a scale of 1-5. Conversely for blue ($x < 0$) cells. Black cells 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.

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, Trujillo, and Fagni 2022). This is likely another artifact caused by alignment procedures, making LLMs too agreeable (Park et al. 2023b).

5.2 Ablation Study

We generate eight synthetic discussions per ablation experiment, using a single model (Qwen 2.5). We compare the diversity (cf. §2.3, §4.3) of these discussions with ones from the CeRI “Regulation Room” dataset[2], which includes moderated online deliberative discussions for ten diverse topics.

Effects of LLMs

[2] 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.

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$\times$time	-0.003
Moderation Game$\times$time	-0.011*
Rules Only$\times$time	-0.008
Regulation Room$\times$time	-0.023***
Constructive Communications$\times$time	-0.023***

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

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

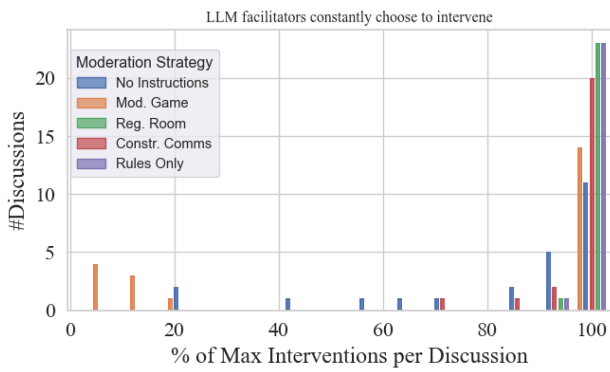


Figure 4: Histogram of interventions by LLM facilitators per strategy used.

Larger models do not lead to more high-quality discussions. As shown in Fig. 6a, 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.

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. 6b 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).

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

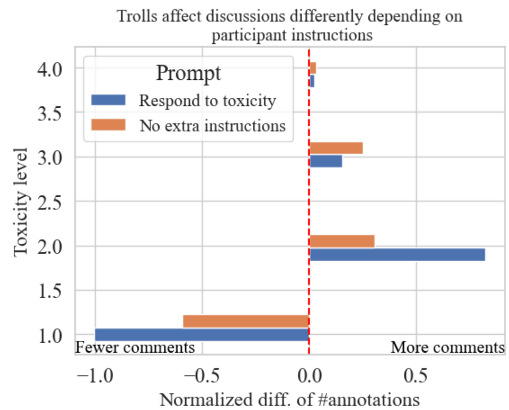


Figure 5: Non-troll toxicity levels in discussions with and without trolls. There is a significant uptick on the number of "somewhat toxic" ($Toxicity = 2$) comments when the participants are primed to respond to toxic comments.

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. 5).

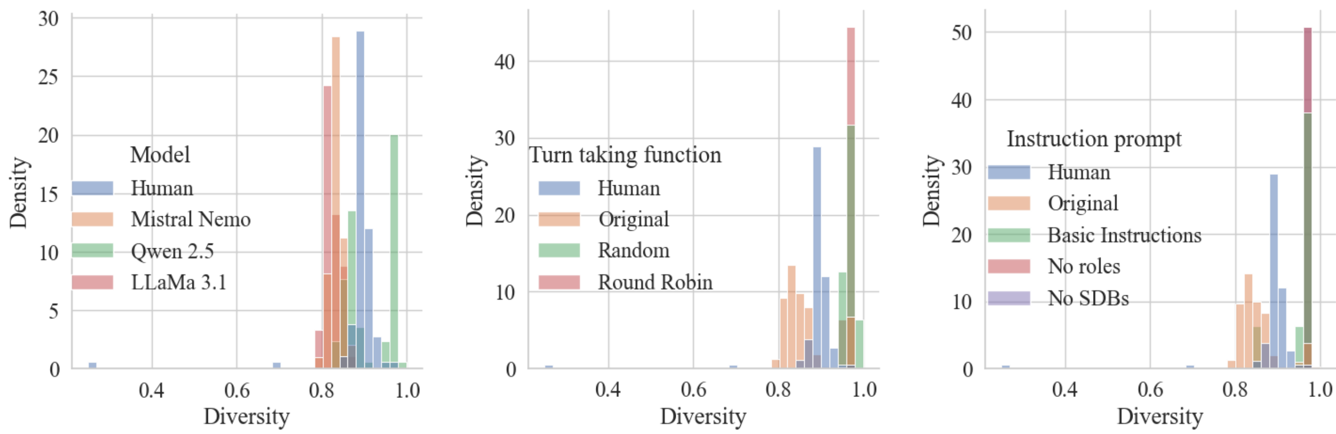
SDBs, roles, and our instruction prompt all increase the quality of synthetic data. Fig. 6c illustrates that incorporating SDBs, roles, and specialized instruction prompts, results in diversity scores more closely aligned with human discussions.

6 Datasets and Software

We introduce 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. 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 this would necessitate filtering out discussions with constant facilitator interventions or low/extremely high diversity scores. However, the data can be scaled accordingly, due to the low computational cost of our methodology.



(a) Model

(b) Turn-taking function t

(c) Prompting function ϕ

Figure 6: Diversity (§2.3) distribution for each discussion by LLM (§4.4), turn-taking function t (§3.2), and prompting function ϕ used (§3.3).

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 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 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.

8 Discussion

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

Limitations Due to limited research in the area, our analysis uses only two quality metrics to gauge discussion quality: diversity and toxicity. Additionally, since our work is one of the first exploring such simulations for online discussions, there is necessarily no strong theoretical support.

While we investigate the impact of facilitation strategies in synthetic discussions, we cannot claim that the behavior of LLM user- and facilitator-agents is representative of human behavior. This claim can be scarcely made in Social Science studies involving LLM subjects (Rossi, Harrison, and Shklovski 2024; Zhou et al. 2024), as discussed in §2.4.

Furthermore, our experimental setup makes several assumptions that may affect the generalizability of our findings. We examine only three LLMs, assume a maximum of one facilitator per discussion, and use a turn-taking algorithm that overlooks contextual factors like relevance and emotional engagement, which are important in human interactions (Rooderkerk and Pauwels 2016; Ziegele et al. 2018). Moreover, due to resource constraints, we were unable to experiment with more elaborate instruction prompts, due to the need for large context windows. Our methodology also does not account for the fact that humans may behave differently when knowing they are interacting with LLMs instead of humans,.

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). Such actors could adapt our methodology to maximize toxicity, disrupt human discussions, or learn to circumvent moderation mechanisms to propagate misinformation or spread specific agendas. Notably, LLMs currently lack robust defenses against these types of attacks (Li et al. 2025), although ongoing research is addressing these vulnerabilities (Wang et al. 2025).

Furthermore, the inherent biases within LLMs risk skewing moderation systems towards the predominant demographics best represented in their training data, often at the expense of disadvantaged or underrepresented groups (Rossi, Harrison, and Shklovski 2024; Anthis et al. 2025; Burton et al. 2024). While the use of SDB prompts is a necessary step toward inclusivity, it remains insufficient for verifiable, equitable representation (Rossi, Harrison, and Shklovski 2024).

Finally, it is crucial to repeat that while LLMs can approximate aspects of human behavior, they do not reliably replicate it (§2.4). Consequently, this research should be viewed as a foundation for pilot experiments, and conclusions about human behavior should be drawn with caution when based solely on synthetic data.

References

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.

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.

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.

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.

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.

OpenAI; Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; Avila, R.; Babuschkin, I.; Balaji, S.; Balcom, V.; Baltescu, P.; Bao, H.; Bavarian, M.; Belgum, J.; Bello, I.; Berdine, J.; Bernadett-Shapiro, G.; Berner, C.; Bogdonoff, L.; Boiko, O.; Boyd, M.; Brakman, A.-L.; Brockman, G.; Brooks, T.; Brundage, M.; Button, K.; Cai, T.; Campbell, R.; Cann, A.; Carey, B.; Carlson, C.; Carmichael, R.; Chan, B.; Chang, C.; Chantzis, F.; Chen, D.; Chen, S.; Chen, R.; Chen, J.; Chen, M.; Chess, B.; Cho, C.; Chu, C.; Chung, H. W.; Cummings, D.; Currier, J.; Dai, Y.; Decareaux, C.; Degry, T.; Deutsch, N.; Deville, D.; Dhar, A.; Dohan, D.; Dowling, S.; Dunning, S.; Ecoffet, A.; Eleti, A.; Eloundou, T.; Farhi, D.; Fedus, L.; Felix, N.; Fishman, S. P.; Forte, J.; Fulford, I.; Gao, L.; Georges, E.; Gibson, C.; Goel, V.; Gogineni, T.; Goh, G.; Gontijo-Lopes, R.; Gordon, J.; Grafstein, M.; Gray, S.; Greene, R.; Gross, J.; Gu, S. S.; Guo, Y.; Hallacy, C.; Han, J.; Harris, J.; He, Y.; Heaton, M.; Heidecke, J.; Hesse, C.; Hickey, A.; Hickey, W.; Hoeschele, P.; Houghton, B.; Hsu, K.; Hu, S.; Hu, X.; Huizinga, J.; Jain, S.; Jain, S.; Jang, J.; Jiang, A.; Jiang, R.; Jin, H.; Jin, D.; Jomoto, S.; Jonn, B.; Jun, H.; Kaftan, T.; Łukasz Kaiser; Kamali, A.; Kanitscheider, I.; Keskar, N. S.; Khan, T.; Kilpatrick, L.; Kim, J. W.; Kim, C.; Kim, Y.; Kirchner, J. H.; Kiros, J.; Knight, M.; Kokotajlo, D.; Łukasz Kondraciuk; Kondrich, A.; Konstantinidis, A.; Kosic, K.; Krueger, G.; Kuo, V.; Lampe, M.; Lan, I.; Lee, T.; Leike, J.; Leung, J.; Levy, D.; Li, C. M.; Lim, R.; Lin, M.; Lin, S.; Litwin, M.; Lopez, T.; Lowe, R.; Lue, P.; Makanju, A.; Malfacini, K.; Manning, S.; Markov, T.; Markovski, Y.; Martin, B.; Mayer, K.; Mayne, A.; McGrew, B.; McKinney, S. M.; McLeavey, C.; McMillan, P.; McNeil, J.; Medina, D.; Mehta, A.; Menick, J.; Metz, L.; Mishchenko, A.; Mishkin, P.; Monaco, V.; Morikawa, E.; Mossing, D.; Mu, T.; Murati, M.;

Murk, O.; Mély, D.; Nair, A.; Nakano, R.; Nayak, R.; Neelakantan, A.; Ngo, R.; Noh, H.; Ouyang, L.; O'Keefe, C.; Pachocki, J.; Paino, A.; Palermo, J.; Pantuliano, A.; Parascandolo, G.; Parish, J.; Parparita, E.; Passos, A.; Pavlov, M.; Peng, A.; Perelman, A.; de Avila Belbute Peres, F.; Petrov, M.; de Oliveira Pinto, H. P.; Michael; Pokorny; Pokrass, M.; Pong, V. H.; Powell, T.; Power, A.; Power, B.; Proehl, E.; Puri, R.; Radford, A.; Rae, J.; Ramesh, A.; Raymond, C.; Real, F.; Rimbach, K.; Ross, C.; Rotsted, B.; Roussez, H.; Ryder, N.; Saltarelli, M.; Sanders, T.; Santurkar, S.; Sastry, G.; Schmidt, H.; Schnurr, D.; Schulman, J.; Selsam, D.; Sheppard, K.; Sherbakov, T.; Shieh, J.; Shoker, S.; Shyam, P.; Sidor, S.; Sigler, E.; Simens, M.; Sitkin, J.; Slama, K.; Sohl, I.; Sokolowsky, B.; Song, Y.; Staudacher, N.; Such, F. P.; Summers, N.; Sutskever, I.; Tang, J.; Tezak, N.; Thompson, M. B.; Tillet, P.; Tootoonchian, A.; Tseng, E.; Tuggle, P.; Turley, N.; Tworek, J.; Uribe, J. F. C.; Vallone, A.; Vijayvergiya, A.; Voss, C.; Wainwright, C.; Wang, J. J.; Wang, A.; Wang, B.; Ward, J.; Wei, J.; Weinmann, C.; Welihinda, A.; Welinder, P.; Weng, J.; Weng, L.; Wiethoff, M.; Willner, D.; Winter, C.; Wolrich, S.; Wong, H.; Workman, L.; Wu, S.; Wu, J.; Wu, M.; Xiao, K.; Xu, T.; Yoo, S.; Yu, K.; Yuan, Q.; Zaremba, W.; Zellers, R.; Zhang, C.; Zhang, M.; Zhao, S.; Zheng, T.; Zhuang, J.; Zhuk, W.; and Zoph, B. 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. *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*.

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. arXiv:2304.03442.

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.; 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.

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.

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 community 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.

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.

Zhou, X.; Su, Z.; Eisape, T.; Kim, H.; and Sap, M. 2024. Is this the real life? Is this just fantasy? The Misleading Success of Simulating Social Interactions With LLMs. In Al-Onaizan, Y.; Bansal, M.; and Chen, Y.-N., eds., *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 21692–21714. Miami, Florida, USA: Association for Computational Linguistics.

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.