Designing Synthetic Discussion Generation Systems: A Case Study for Online Facilitation

DIMITRIS TSIRMPAS, ION ANDROUTSOPOULOS, and JOHN PAVLOPOULOS, Athens University of Economics and Business, Greece and Archimedes, Athena Research Center, Greece

arXiv:2503.16505v4 [cs.HC] 17 Apr 2026

A critical challenge in social science research is the high cost associated with experiments involving human participants. While some studies have explored the use of Large Language Models (LLMs) as imperfect substitutes, there has been little work on how to define, design, and evaluate such experiments. In this paper, we identify Synthetic Discussion Generation (SDG), a novel Natural Language Processing (NLP) direction aimed at creating simulated discussions that enable cost-effective pilot experiments. Drawing on existing SDG systems and interdisciplinary literature, we develop a theoretical, task-agnostic framework for designing, evaluating, and implementing these simulations. We argue that the use of proprietary models such as the OpenAI GPT family for such experiments is often unjustified in terms of both cost and capability, despite its prevalence in current research. Our experiments demonstrate that smaller quantized models (7B–8B) can produce effective simulations at a cost more than 44 times lower compared to their proprietary counterparts. We use our framework in the context of online facilitation, where humans actively engage in discussions to improve them, unlike more conventional content moderation, which only removes inappropriate content from discussions. The extremely large scale of modern social networks has led researchers to develop LLM facilitators, whose capabilities remain largely unassessed due to the need for costly experiments with human discussants. By treating this problem as a downstream task for our framework, we show that synthetic simulations can yield generalizable results at least by revealing limitations before engaging human discussants. In LLM facilitators, a critical limitation is that they are unable to determine when to intervene in a discussion, leading to undesirable frequent interventions and, consequently, derailment patterns similar to those observed in human interactions. Additionally, we find that different facilitation strategies influence conversational dynamics to some extent. Beyond our theoretical SDG framework, we also present a cost-comparison methodology for experimental design, an exploration of available models and algorithms, an open-source Python framework, and a large, publicly available dataset of LLM-generated discussions across multiple models.

CCS Concepts: • **Computing methodologies** → **Natural language generation**; *Discourse, dialogue and pragmatics*; *Language resources*; • **Human-centered computing** → *Open source software*; Social networking sites; Social media; Collaborative and social computing design and evaluation methods.

Additional Key Words and Phrases: LLMs, discussions, facilitation, synthetic, experiments

ACM Reference Format:
Dimitris Tsirmpas, Ion Androutsopoulos, and John Pavlopoulos. 2026. Designing Synthetic Discussion Generation Systems: A Case Study for Online Facilitation. *ACM Trans. Soc. Comput.* 1, 1 (April 2026), 46 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

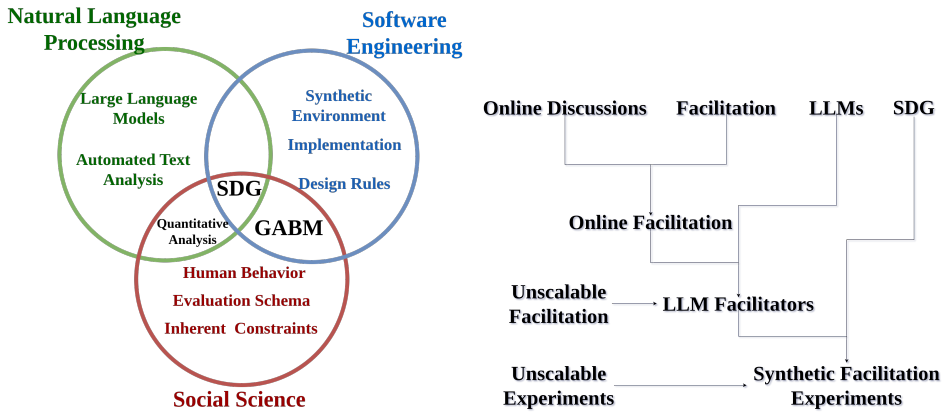

Fig. 1. **Left**: Synthetic Discussion Generation (SDG) is a subset of Generative Agent-Based Modelling (GABM) focused exclusively on interactions through text. Designing SDG systems requires knowledge from three disciplines: social science helps us understand how the system should function, Natural Language Processing (NLP) provides the tools to automate generation and evaluation, and Software Engineering (SWE) enables us to build a scalable and generalizable system. **Right**: We can use SDG to solve to discover and debug issues in the emerging field of LLM facilitation, without the need for costly human experiments.

1 Introduction

Advances in generative Artificial Intelligence (AI) technologies have enabled the study of human behaviors through Large Language Models (LLMs). Since these models are pretrained on extensive human discussion data, researchers have hypothesized that they could replicate certain human behaviors in social science experiments concerning behavioral social science studies [7, 39, 115] or human-algorithm interaction [21, 52, 77], usually by creating synthetic simulation platforms [10, 22, 57, 77, 87, 115, 128]. However, since the emergent properties of LLMs have only been discovered relatively recently, studies using synthetic agents so far have employed various ad-hoc methodologies to define, run, and evaluate simulated discussion experiments [1, 67, 76–78, 115, 116]. In an attempt to consolidate the currently fragmented research landscape of synthetic simulations, we are the first to identify and define this, so far implicit, research direction:

Definition. **Synthetic Discussion Generation (SDG)**: Simulating textual human discussions, by using synthetic participants, ultimately aiming to understand and extract representative insights for human discussions.

Authors' Contact Information: Dimitris Tsirmpas, dim.tsirmpas@aueb.gr; Ion Androutsopoulos, ion@aueb.gr; John Pavlopoulos, annis@aueb.gr, Athens University of Economics and Business, Informatics, Athens, Greece and Archimedes, Athena Research Center, Archimedes Unit, Athens, Greece.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
© 2026 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM 2469-7826/2026/4-ART
https://doi.org/10.1145/nnnnnnn.nnnnnnn

SDG is a specific case of Generative Agent-Based Modelling (GABM), which is a computational approach that simulates the interactions of agents to study complex behaviors and emergent phenomena. Unlike generalized GABM systems, SDG actors do not have a predetermined set of actions and behaviors; rather, both are expressed through free text. SDG is typically used as a means to solve research questions in other tasks (called *'downstream tasks'* in this study) where both human participation and a sustained conversational context are needed, as opposed to single exchanges (in which case, simple prompting would be sufficient). Examples include negotiations [1], social experiments [77] and human persona replication [78]. Despite the breadth of downstream tasks, we argue that a well-designed SDG methodology can generalize over most such tasks.

The lack of a unified design methodology in studies using SDG results in greater difficulty in creating, running and assessing the validity of these experiments. Furthermore, *there has been no study that addresses how to systematically create, evaluate, and use generalizable LLM simulations of discussions to solve research questions* without the need for human participants and evaluators. In this study, we answer the following research question (RQ):

RQ. How do we design and evaluate scalable[1] SDG simulations that (a) replicate known representative behaviors found in real-life human discussions, and (b) can be used to obtain further insights on real-world phenomena?

We begin by creating a theoretical, downstream-task-agnostic framework for SDG. Since there are no standards or guidelines on how to create SDG systems, we examine task-specific systems in the literature and use lessons from Software Engineering (SWE) to derive design rules (§3.1). Through this analysis, we formalize Synthetic Discussion Generation (SDG) as a distinct research direction and identify four core design principles for effective synthetic simulations. We break down the requirements for SDG and derive a set of isolated components that must be implemented and separately evaluated (§3.3, §3.4), including the construction of LLM personas and the execution of dynamic discussions. To assess whether observed behaviors are representative of real-world patterns, we build an evaluation suite grounded in GABM literature and adapted for SDG, which we use to verify the necessity and integrity of these components (§3.2). Our findings show that persona-driven simulations improve the diversity of generated content, aligning it more closely with human discussions, while careful instruction prompting can induce persistently toxic behavior in safety-aligned LLMs using only conversational context. Additionally, we observe that LLM agents tend not to abstain from participation, showcasing an inherent limitation in modeling realistic conversational dynamics.

Having created a theoretical SDG framework, we tackle the issue of scalability: how can we minimize the required cost for running simulated experiments? Most recent studies [1, 10, 76, 77, 115] use proprietary models such as the OpenAI GPT family. We argue that the use of such models increases inference costs, which inherently limits the amount, efficiency, and scalability of experiments conducted using these models in real-world problems (§4.1). Indeed, we show that small (7B-8B parameter) quantized models are capable enough for our purposes (Finding 1) and able to replicate several social dynamics observed in human discussions (Appendix B) at a fraction of the cost. By developing a methodology for cost comparison between experiments with humans, proprietary models, and locally hosted LLMs (§4.1), we estimate that our approach using smaller models achieves a *cost reduction of 1,600 times* compared to experiments with humans and *44 times* compared to a recent, mid-capability proprietary model (GPT-5.1–see Table 8).

Next, we investigate whether such simulations can produce insights transferrable to human discussions, by using LLM facilitation as a downstream task (§5). Platform designers and researchers

[1]Scalable in this context implies that the system can maintain consistent performance and quality as the scale of the experiments increases, without requiring disproportionate increases in computational resources or financial cost.

have traditionally focused on identifying, flagging, and removing problematic content, an approach often called "content moderation" in literature [26, 98]. However, fully automatic content moderation is insufficient in practice [44, 52, 93, 99]. A more effective approach is using human facilitators,[2] who actively participate in a discussion. Still, this approach cannot be used scalably in modern social media sites where hundreds of thousands of discussions are being conducted each day. LLMs have been proposed as scalable facilitators [52, 99], yet experimentation and development remain costly and constrained by the need for human discussants [89, 99]. The main issue is that facilitation cannot be meaningfully evaluated in isolated, single-turn listen-and-respond settings; it requires sustained conversational context [21], with multiple human participants actively conversing with the LLM facilitator. Our study is the first to use multiple participants to evaluate a LLM facilitator, and the first to use synthetic participants in lieu of humans.

While the use of LLM participants and the relatively limited recent work on facilitation constrain the scope of our findings, our experiments uncover generalizable results. Specifically, we identify a severe and previously unreported limitation: LLM facilitators cannot choose not to intervene, even when intervention is unnecessary (Finding 2). This inability to remain silent has not, to our knowledge, been documented in prior literature. This behavior has been observed to induce frustration and toxicity among human participants in online communities [52]. [3] Most importantly, failing to uncover this issue could lead researchers to invest significant time and resources in human-subject studies and complex LLM configurations [52, 99], only to find their results fundamentally undermined. We also identify two real-life facilitation strategies and two common setups used in current LLM facilitation systems and through our experiments confirm that LLM facilitators are able to alter conversational dynamics according to different facilitation strategies (Finding 3).

Finally, we release an open-source Python framework[4] that implements our methodology at scale (§6.1), enabling the research community to rapidly experiment with synthetically generated discussions. Given that existing facilitation datasets are few and generally small [52], we also release a large, publicly available dataset with LLM-generated synthetic discussions (§6.2). Our dataset can be used for further analysis of the behavior of out-of-the-box LLMs in synthetic discussions generally, and as LLM facilitators specifically, as well as for LLM facilitator fine-tuning [116] in future work. We use open-source LLMs and include all relevant configurations in order to make our study as reproducible as possible.

2 Related Work

2.1 Substituting human subjects with LLMs

SDG can be viewed as a specialization of GABM, which uses generative agents to simulate believable human behavior [77]. Early approaches to such simulations relied on rules-based and Machine Learning (ML)-driven models [28], and have successfully demonstrated the emergence and replication of complex social dynamics across a range of scenarios [86, 94]. In this context, LLMs can similarly act as agents within controlled simulated environments, extending the capabilities of prior approaches. LLMs can naturally exhibit bias and forms of irrationality in their behavior, which constituted a fundamental limitation of earlier models that tended to enforce overly rational decision-making [2].

SDG systems include synthetic clones of Reddit [77], Twitter/X [67], generic social media sites [87, 115], games [76], and social experiments [22, 58, 128]. One approach [10] extracts topics and comments from online human discussions and prompts an LLM to recreate them based on

[2]Some publications use the term "moderator" with the meaning we have assigned to "facilitator" [52].
[3]A pattern we replicate in our experiments (App. B; Table 14).
[4]Website: https://dimits-ts.github.io/syndisco, Code: https://github.com/dimits-ts/syndisco

summaries. Another approach [116] creates synthetic discussions between two roles: an agent controlling a fictional environment and another client-agent interacting with it. These discussions are filtered and used to finetune the agent LLM for a specific task. Finally, a third approach [1] generates synthetic negotiations involving multiple agents with different agendas and responsibilities. Our experiments can be viewed as a domain shift from negotiation to discussion facilitation, where various user types (e.g., normal users, trolls) engage in discussion overseen by an actor with veto power (in our case, the facilitator).

Recent work [7, 39, 115] argues that synthetic agents have the potential to eventually replace human participants in social science studies. Indeed, LLMs have demonstrated complex, emergent social behaviors. In a simulated setting, LLM actors have been successful in diffusing information through social interaction, as well as planning and organizing spontaneous social events without human interaction [76]. In settings where LLM actors represent distinct factions with goals and agendas, they were shown to execute basic social strategies and adapt their plans in response to the other LLM actors [1]. Additionally, substituting SocioDemographic Backgrounds (SDBs) with information from personal interviews enables LLMs to reliably predict further survey responses from interviewed humans [78].

However, significant limitations of LLMs remain in the context of Social Science experiments. Issues include undetectable behavioral hallucinations [89]; sociodemographic, statistical and political biases [5, 42, 89, 108]; unreliable annotations [13, 37, 48, 69]; non-deterministic outputs [8, 13]; and excessive agreeableness [5, 76, 89]. Even if LLMs can mimic some human behaviors, they are ultimately at most a statistical estimate of the "average" human encountered in their training data [89]; thus we can not claim that any LLM behavior can be indicative of human behavioral patterns without extensive replication studies [89] or system evaluation [2]. This is a core, unavoidable limitation of any LLM simulation. Rather than asserting that the findings from our experiments will be directly replicable in human studies, we focus on reporting results that are likely to be applicable to real-world settings, based on thorough evaluation.

2.2 Evaluation

2.2.1 Synthetic data quality. When creating SDGs we need to evaluate *synthetic data quality*, as discussions without human involvement tend to deteriorate quickly, often becoming repetitive and low-quality [116]. Despite the need for such metrics, methods for quantifying the quality of synthetic data remain limited [116].

One approach uses a mix of graph-based and lexical similarity metrics, many of which depend on human discussion datasets for comparison [10]. The authors also propose a loosely defined "coherence" score, which is LLM-annotated without theoretical grounding. Another approach assesses quality through post-discussion surveys with humans, and by measuring lexical diversity to approximate the variety of opinions expressed [50].

Diversity [116] is a metric applied to discussions that penalizes repeated text sequences between comments[5] of the same discussion using average pairwise ROUGE-L[6] [59] scores. Low diversity points to pathological problems with model generation (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). Specifically, diversity is defined as follows, where N_d is the length (in comments) of discussion d, and $c(i, d)$ is the ith comment of

[5]Also called posts or, in older publications, dialogue turns.

[6]ROUGE-L measures the similarity between two texts by identifying the longest sequence of words that appear in both in the same order.

discussion d.

$$div(d) = 1 - \frac{2}{N_d(N_d - 1)} \sum_{i=1}^{N_d-1} \sum_{j=i+1}^{N_d} ROUGE\text{-}L(c(i,d), c(j,d))$$

This metric inherits some limitations of ROUGE scores, but it remains computationally efficient, interpretable, and domain-agnostic. We argue that the use of exact-word-matching, inherited from ROUGE, is not necessarily a disadvantage in SDG. In fact, this approach provides the most direct and reliable way to detect "collapsed discussions" or model repetitions [116]. Another advantage of this approach in SDG is that humans frequently repeat specific phrases or quote exact portions of prior dialogue ('other-repetition') to respond to particular points, signal understanding, or summarize previous exchanges [15, 88, 107]. Additionally, alternative metrics such as BLEU and BERTScore suffer from significant limitations in both practical applicability and reproducibility, despite the latter not relying on exact-word-matching [20, 41, 81]. For these reasons, we incorporate diversity in our analysis while acknowledging that more robust synthetic quality measures are still required.

2.2.2 GABM evaluation. When evaluating SDG systems, a naive assumption would be to test whether the text itself is "believable", which could be evaluated by tasking humans with discriminating between synthetic and human-written discussions [77]. This approach is insufficient, since LLMs are explicitly trained to convincingly mimic human text, and because believable human text does not translate to believable social behaviors [89]. We therefore turn to GABM literature to find robust validation methods for SDG systems.

There is currently no consensus on how a GABM system should be evaluated [2]. We can use automated methods developed for *non-generative* agent-based models (foundational and empirical methods [25]), or humans to judge observed behaviors and dynamics (face validation [85]). Since the original procedures were designed with ML and rule-based models in mind, a recent study [2] proposes an adapted set of guidelines for generative agents, which we briefly outline below.

Foundational and Empirical Methods. Foundational methods ensure data integrity and consistency with established benchmarks, without requiring manual evaluation:

- **Data analytics**: A prerequisite for all further analysis. Includes rigorous cleaning, organization, and statistical analysis to ensure data integrity.
- **Docking**: Also known as 'model-to-model comparison', this involves comparing a simulation's outputs with those of other state-of-the-art ABMs to verify alignment. It is particularly useful when a recognized reference model exists.
- **Empirical validation**: This strategy uses real-world historical or observational data to ensure that model predictions are grounded in reality. It is the primary means of evaluating a model's accuracy in replicating known phenomena.
- **Sampling**: Sampling involves exploring the simulation space using various techniques to understand how changes in input variables (e.g., decision thresholds) impact outcomes.
- **Visualization**: This family of methods utilizes graphical representations to analyze emergent behaviors and patterns. While powerful for detecting anomalies, it is subject to qualitative interpretation biases.

Face-Validation Methods. Face-validation focuses on whether model processes and outcomes appear "reasonable" to humans. It is especially critical for GABMs, where natural language and social dynamics are central. Unlike empirical methods, face-validation does not strictly require external datasets, but does require manual labor.

- **Survey-based validation**: Human evaluators, often dozens or hundreds, are engaged to rank or score agent behaviors based on their perceived believability and social intelligence.

An alternative interpretation suggests that the agents themselves could be "interviewed", and their responses be judged by humans [2].

- **Observational studies**: These methods analyze human behavior in real-world settings to compare them with generative agents. An example would be a study evaluating synthetic social media simulations by comparing generated responses to documented posts during a real-life disaster in Japan [35]. While empirical validation (see above) uses aggregate statistics from humans, these methods analyze behaviors on a higher level of abstraction.
- **Expert reviews**: Domain experts conduct structured walkthroughs or immersive assessments where they may even adopt the perspective of an agent to directly evaluate the coherence and plausibility of interactions. This approach has been used in work relating to the effectiveness of LLM facilitators [21], but this was made possible by the particular setup used, the relatively low number of samples, and paying participants relatively low wages.

2.3 Online Facilitation

2.3.1 Human Facilitation. The modern social media environment has evolved to be extremely hostile, with users facing ever-increasing threats such as targeted misinformation [23, 27], hate speech [51], and polarization [84]. These threats can cause serious emotional and mental harm [95], radicalization [21], real-world violence [93], they may sabotage democratic dialogue [30, 32, 98], and undermine trust in democratic institutions [97]. Threats of these kinds have been greatly amplified with the advent of LLMs; malicious actors can use generative AI technology to mass-produce misinformation [53, 105], attack online infrastructure [105], and overwhelm online communities with synthetic content [92]. On the other hand, evidence shows that human moderators experience elevated levels of psychological distress, including secondary trauma [103], compassion fatigue, vicarious trauma and burnout [102], as well as exhaustion [104], with symptoms worsening through prolonged exposure [103].

In light of the limitations of traditional content moderation, researchers have increasingly turned to approaches in which moderators actively engage with users through dialogue–often termed conversational moderation or facilitation [7, 21, 32, 52]. Rather than reacting to violations after they occur, facilitation aims to preempt problematic behavior [3, 21, 26, 98] while fostering deliberation and collective decision-making within communities [50, 98]. Human facilitators play a crucial role in online communities, serving as key contributors to maintaining social cohesion, encouraging participation, and guiding discussions in a constructive manner [4]. Their role is often informal and community-driven, with facilitators typically selected by the community members themselves [4] and functioning as de-facto community leaders [98]. As such, they are not only moderators in the traditional sense but also volunteers who embody the values and norms of the group [52]. The diversity of approaches to human moderation reflects the complexity of online interaction, with various strategies employed to manage discourse, resolve conflicts, and ensure equal participation [18, 36, 75, 97].

2.3.2 LLM Facilitators. Human facilitators face many of the same psychological challenges as content moderators [96], and cannot handle the sheer scale of modern-day social network traffic [93], limiting the scalability of purely human-driven facilitation approaches. Given the scale of threats to online communities and the growing capacity to mass-produce harmful content, reliance on a predominantly human-centered moderation model is increasingly untenable; we must thus turn to automated tools [52, 99]. Unlike classification models often used in online platforms to detect content violating norms, LLMs can actively facilitate discussions [52]. They can warn users for rule violations [55], monitor engagement [97], aggregate diverse opinions [99], and provide translations

and writing tips, which are especially useful for marginalized groups [111]. These capabilities suggest that LLMs may be able to assist or even replace human facilitators in many tasks [52, 99].

Despite the potential of LLMs in this role, very few studies have actually implemented and evaluated LLM facilitators. One approach used simple rule-based models [50] and reported promising results in organizing user activity. Another study used LLM facilitators to privately discuss with a single human who pretended to be a participant in a by-then-archived discussion [21]. In contrast to both works, our experiments use multiple exclusively LLM participants and LLM facilitators, evaluating the latter within an explicitly toxic and challenging environment. A third approach uses LLM participants, trolls, and moderators with SDBs in a synthetic clone of Reddit [77]. While our experimental setups are similar, our study (1) attempts to create a generalizable framework instead of a task-specific experimental setup, (2) uses local models instead of proprietary models used in the study, (3) provides a robust and scalable evaluation setup instead of human annotators discriminating between synthetic and human-authored text (see §2.2 for why this evaluation setup is insufficient, and §3.2 for scalability concerns).

3 Designing robust SDG frameworks

3.1 Design

3.1.1 Scope. We study SDG in the context of text-only, online discussions. Unlike some online fora and social media such as Reddit, which feature tree-style discussions ('threads') [52], we assume a chat-like schema often encountered in comment sections or IRC chats, where a global ordering exists between all comments in a discussion. This greatly simplifies the design of our framework, although a small modification to how context is handled (§3.4.3) would allow our framework to accommodate discussion threads too.

3.1.2 Requirements. Although SDG is widely employed as a research mechanism, relatively few general-purpose tools have been developed to support researchers in generating and managing simulated discussions. Concordia [118] is among the few generalized frameworks explicitly designed to orchestrate discussions between LLM agents. To our knowledge, however, it has been used in only one empirical study [68] and one competition [100] since its release. In contrast, most existing work relies on ad hoc, purpose-built frameworks (see §2.1). Given the rapid growth of the SDG ecosystem following the widespread adoption of LLMs [2], the limited uptake of standardized tooling is striking. Furthermore, despite numerous implementations, the factors that shape–or hinder–broader community adoption of a SDG framework remain underexplored in the literature. To address this gap, we first outline plausible reasons for this limited adoption and then derive corresponding design principles to guide future framework development.

Principle 1 (Simplicity). **The framework should remain as simple as possible**. In the field of SWE, there is a widely accepted principle that simpler systems generally perform their intended functions more effectively: the KISS principle ("Keep It Simple, Stupid") [12, 110]. Empirical evidence from real-world applications supports this notion [11, 71]. Violating this principle may partly explain the difficulties encountered in other SDG systems [10]. For example, complex and inflexible architectures, as well as high initial computational requirements and operating costs, have been identified as limiting factors in systems such as Concordia [100]. Simplicity has likewise been highlighted as a core design principle for online deliberative platforms like BCAUSE [4], which our framework partly simulates.

Principle 2 (Least-bias). **The design decisions should be as unbiased as possible**. Each new component in a SDG framework may implicitly encode researcher bias, as it needs to implement

dynamics that are implicit in the real world (e.g., turn-taking–§3.4.2). Recent work on LLM behaviors [76] managed to derive interesting insights (§2.1) exactly because the authors only provided a simulated environment *without* attempting to encode specific behaviors and interactions in the LLM users; an observation generally supported by GABM literature [2].

Principle 3 (Scalability). **SDG should be scalable.** The primary reason for designing and using SDG methodologies is to bypass the monetary and time cost of human experiments [89]. A large amount of discussions allows us to filter out low-quality experiments [116], enable robust statistical modeling, investigate multiple experimental configurations, and investigate events that are rare in simulations. This is one of the most important design principles, influencing evaluation (§3.2), experiment implementation (§4.1), model selection (§4.2), and constraining the tools with which we can influence LLM participants (Principle 4). A direct consequence of Principle 3 is that a single model should ideally be used for all generation in SDG experiments. We discuss the drawbacks of using multiple such models in detail in §4.1, which lead us to Principle 4.

Principle 4 (Default-model). **Updating any model internals (e.g., finetuning) should be used as a last resort**. Besides the computational cost of modifying the model itself, we would need to load additional LLMs into VRAM memory (since finetuning means modifying the model weights for only some of the participants), multiplying the capital costs associated with the experiments (§4.1.1), and violating Principle 3.[7] Additionally, doing so would inject a degree of bias in our models [108], violating Principle 2.

Absolute adherence to these principles may not be optimal. For example, due to LLM alignment procedures, simple prompting is not enough to induce toxic behavior–a prerequisite for our downstream task. Inserting our own prompts to induce this behavior violates Principe 2, but is necessary for our experiments, as discussed in §3.3.3. We outline a simple documentation system for explaining the justification of each component in Appendix E.

3.2 Evaluation

In §2.2.2, we outlined existing evaluation methods for GABM systems. Table 1 examines which of these methods are applicable to SDG systems, as well as which are suitable for evaluating our facilitation experiments. These methods are often used in combination within the evaluation pipeline. For instance, we manipulate participant roles and instruction prompts (*sampling*)[8] to assess the impact of toxic agents (§3.3.2,§3.3.3), by comparing aggregate statistics with human discussions (*empirical*), examining emergent behaviors (*observation*), as well as *visualizing* and testing for statistically significant effects (*data analytics*) on participant behavior. To ensure scalability, the framework does not rely on human-in-the-loop validation.

To measure how content variety changes across discussions, we compute a *div* score for each discussion (§2.2.1). For each experimental configuration, we aggregate the *div* scores from all discussions generated under that configuration to form an empirical distribution. Because our dataset includes a large number of discussions (Table 12), we treat these distributions as representative. We then compare the distribution from a real-world human dataset (Appendix C.4) with those from our experimental configurations using Jensen–Shannon divergence–an instance of empirical validation. Lower divergence indicates that the experimental discussions more closely match the diversity observed in human discussions. See Appendix C.3 for further details. Since a core goal of our simulations is scalability, we can not use any methods relying on human validation, however.

[7]Proprietary, API-based models bypass the increase in capital expenses, but may charge higher prices for using customized models [73].

[8]The term is notably different from the statistical use of the term, and refers to use of ablation experiments (see §2.2.2).

Table 1. Summary of GABM evaluation strategies and their practical feasibility for SDG. ✓indicates application; × indicates exclusion. For each strategy, we detail its use within our work or provide a justification for its omission.

Name	SDG	Use / Reason for non-use
Data analytics	✓	Data cleaning and post-processing throughout. Statistical analysis for trends in toxicity (§5.1, §5.2).
Docking	×	No established state-of-the-art GABM for SDG exists to serve as a comparison benchmark.
Empirical validation	✓	We compare content variety between human and synthetic discussions (§3). However, limited real-world data on turn taking, and generally facilitation, restricts the extent of our findings.
Sampling	✓	We explore the effects of each isolated component by conducting extensive ablation studies for LLM personas (§3.3.1), roles (§3.3.2), instruction prompts (§3.3.3), seed opinions (§3.4.1) and turn-taking (§3.4.2).
Visualization	✓	Used to verify effective adversarial behavior from LLM users (§3.3.2), and to showcase aberrant behavior exhibited by LLM facilitators (§5.1).
Observational studies	✓	We rely on extensive literature documenting human behavior in online spaces to verify the existence of some emergent social behaviors (Appendix B).
Survey-based validation	×	Requires large numbers of human annotators, as well as domain experts for manual validation. While most such validation techniques are technically supported by our framework, the extensive use of human labor renders all such methods unscalable for SDG.
Expert reviews	×	See above.

In general, a GABM system should be validated based on its purposes and goals, as well as the availability of relevant data [2, 25]. The selection of a subset of validation methods for *inferential* agents (agents that do not learn during the simulation) based on the system's goals and limitations is often used in literature [2]. We therefore consider the proposed evaluation pipeline to be sufficient for a low-stakes, research-oriented system. However, it should be extended for higher-stakes applications, such as production systems that directly interact with users.

3.3 LLM personas

3.3.1 Injecting sociodemographic information. It is well known that LLMs tend to align more closely with WEIRD (Western, Educated, Industrialized, Rich, and Democratic) and socially prominent groups, while diverging from minorities [34, 38, 91] due to their nature as statistical models that adopt the bias of their training data during pretraining. These biases are further shaped by alignment procedures and prompt sensitivity [61, 90]. Thus, synthetic discussions with out-of-the-box LLMs typically result in similar opinions and arguments, with little variance.

Including SocioDemographic Backgrounds (SDBs) (e.g., gender, age and education) in prompts has proven promising in the generation of varied content [14, 101]. Injecting LLM prompts with these personas systematically shifts the distribution of output tokens, hence producing content of larger variety [61]. The creation of these personas is sometimes implemented by randomly selecting SDB attributes from a number of dimensions such as age, gender, and employment [2] or using prompting [77] (which significantly raises inference costs, violating Principle 3). We instead pick

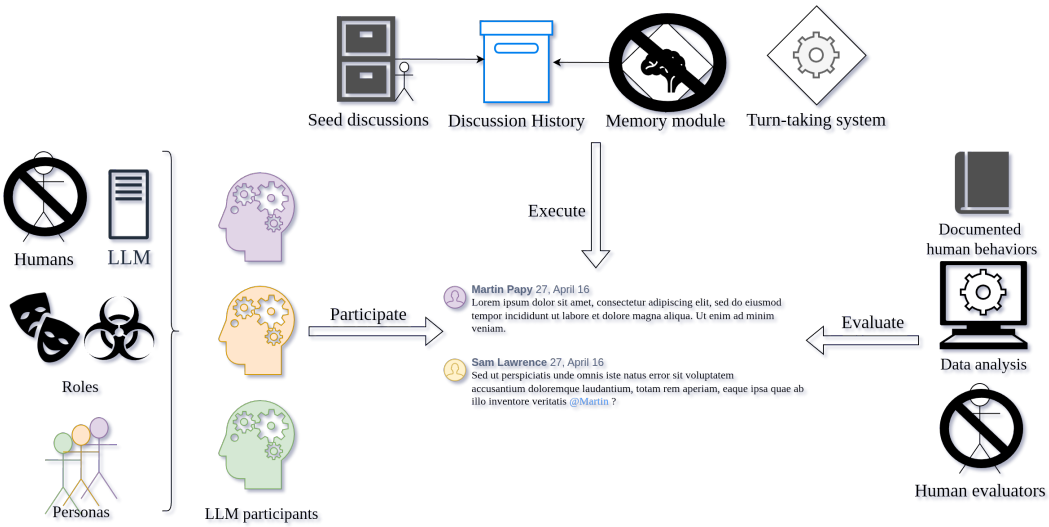

Fig. 2. We identify three key requirements for SDG systems: (1) the generation of LLM participants to replace human agents (§3.3-left), (2) the development of simple synthetic systems capable of managing conversational dynamics (§3.4-top), and (3) the implementation of scalable evaluation methodologies rooted in existing literature, minimizing reliance on human validation (§3.2-right). Note the explicit exclusion of (1) humans for discussion generation and validation, and (2) embedding-based or natural language memory modules used by some simulated experiments.

personas from the PERSONA dataset [16], which includes a curated set of personas based on US census data (Appendix F.1).

Injecting these SDB prompts, we observe that LLM users are able to create and share personal narratives and experiences from the provided information (App. B; Table 13). Additionally, Table 2 shows that using SDBs brings synthetic discussions closer to human-like diversity. Overall, using SDBs seems to be a net positive for our experiments; it is simple (Principle 1), not based on assumptions (Principle 2), inexpensive (Principle 3), explainable, leads to desirable outcomes, and improves content variety. These observations do not necessarily translate to LLM persona behavior indicative of real humans (and indeed relevant work suggests otherwise, see §2.1), but this is not a necessary condition for validating the findings of the experiments on our down-stream task (§5). Specifically, Finding 1 concerns targeted outputs (toxicity, talking patterns) of LLMs, while Findings 2,3 refer to general patterns in LLM facilitator generation without assuming that the LLM participants are "realistic".

Table 2. Jensen-Shannon divergence between the *div* distributions of human and synthetic discussions by SDBs. Smaller is better, lowest values highlighted in bold.

SDBs	LLaMa-70B	LLaMa-8B	Mistral-24B	Mistral-7B	Qwen-32B	Qwen-7B
None	0.395	0.492	0.329	0.489	0.152	0.232
Injected	**0.356**	**0.443**	**0.321**	**0.443**	**0.100**	**0.196**

3.3.2 Adding antagonistic users. The most basic discussion setup for our experiments involves a single role: that of participants who discuss a topic among themselves. Since our end-goal is to investigate the behavior of LLM facilitators, we need to develop a challenging discussion environment. We can assume that the presence of toxicity is a good proxy for a challenging environment, since it has been shown to inhibit discussions by fermenting mistrust [4], inhibiting productive discussions [9] and causing participants to leave [4]. For these reasons, toxicity has been a long-standing target of social science research on online communities [3, 40, 83, 93, 120].

LLMs are inherently aligned to be highly agreeable, making it difficult for them to replicate toxic or harmful behaviors [5]. While fine-tuning can alter this behavior [108], we consider it an inadvisable method for SDG (see Principle 4). An alternative is to use base-pretrained LLMs, but these lack instruction-following capabilities, as no safety-unaligned, instruction-tuned models are available. We are thus left with one choice; attempting to subtly 'jailbreak' LLMs using prompting. LLM jailbreaking refers to the manipulation of an LLM to bypass its built-in safety constraints and generate outputs that would otherwise be restricted or censored. We note that our approach is different from conventional prompt-based jailbreak methods which attempt to optimize a single prompt designed to "fool" the model [123]. Instead of relying on prompts that break model safety by using fictional scenarios, shifting the attention of the model, or escalating privileges [60], we rely on a partially derailed conversational context to induce prohibited behaviors.[9]

Specifically, we assign antagonistic motivations to a randomly selected subset of participants, with a set probability (30%) of each participant being selected for this role. These antagonistic participants are inspired by real-life internet *trolls* who deliberately use provocative, inflammatory, or misleading content to incite conflict, disrupt discussions, or manipulate online interactions for personal gain or amusement.[10] It is worth noting that the rest of the LLM participants are not aware of which users are trolls; they have to infer it from context. This setup is similar to negotiation [1] and social games simulations [70], where participants had to guess which factions their opponents secretly belonged to.

A simple prompt instructing the LLMs to act as trolls is enough to provoke toxic behavior (Fig. 3), following human trends (Fig 7). While their speech may sometimes be stereotypical and unsubtle (validating relevant work suggesting that LLMs often adopt exaggerated, stereotypical features instead of nuanced views of human behavior [105]), the inclusion of trolls leads to emergent social dynamics in our simulations. [11] Therefore, while including explicit troll users violates Principle 2, it remains one of the simplest (Principle 1), least expensive (Principle 3) ways of creating emergent dynamics in our experiments.

3.3.3 Ensuring meaningful interaction. One might expect that the mere presence of trolls in synthetic discussions would prompt other participants to engage with antagonistic speech. However, this is not observed: Fig. 4 (yellow, triangle markers) shows that basic prompts instructing LLMs to behave as humans are insufficient to elicit toxic responses, even in the presence of trolls—an essential prerequisite for subsequent experiments. This limitation is likely due to alignment procedures that discourage models from engaging with toxic content [76, 89].

While finetuning models to increase toxicity could address this issue, it is unsuitable for SDG (see Principle 4). Instead, we find that simple instruction-based interventions are effective: prompting

[9]It is worth noting that very early work on such systems did not have this limitation, since early safety-alignment technology allowed LLMs to be much more likely to produce toxic speech–even using proprietary models [77].

[10]There is no formal definition of what a troll is [64], but we do not need one in this work; LLMs are extensively trained on online corpora and should thus replicate text that is associated with the human perception of the term.

[11]For instance, we observe a transphobic user attempting to use their credentials as a doctor to justify their position (App B; Table 13), a troll being protected by the wider community due to perceived facilitator overreach (Table 15), or becoming the target of collective action from other participants (Table 16).

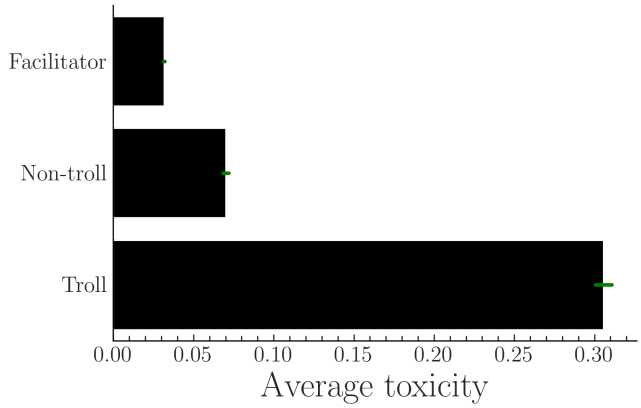

Fig. 3. Average toxicity of non-troll, troll, and facilitator users across synthetic discussions. Our "troll" prompting is necessary for discussions to exhibit human-like toxicity (Fig 7). Error bars indicate 95% confidence intervals. We use the Perspective API [80] to obtain the toxicity scores; see §4.3 for experimental details.

participants to respond to toxic content (the 'Respond-Provoke' strategy; Appendix F) enables the emergence and spread of toxicity, even in the absence of explicit trolls (Fig. 4, first x-axis tick). Moreover, only under this strategy, the toxicity of non-troll participants increases significantly as the number of trolls grows (black, round markers).

We had initially hypothesized that a potential downside of this approach would be that users are biased towards always going against what they perceive as trolling behavior due to our explicit instructions to do so; however, there exist cases where the users back up trolls against the facilitator's judgment, as evidenced by Table 15. This finding suggests that our 'Respond-Provoke strategy' does not significantly compromise Principle 2 and remains one of the simplest (Principle 1), and least expensive (Principle 3) ways of achieving the emergent social properties mentioned in the previous section (§3.3.2). Unfortunately, our strategy does lead to greater divergence from natural human conversations (Table 3).

Table 3. Jensen-Shannon divergence between the *div* distributions of human and synthetic discussions depending on whether we prompted the models to respond to toxicity. Smaller is better, lowest values highlighted in bold.

Instructions	LLaMa-70B	LLaMa-8B	Mistral-24B	Mistral-7B	Qwen-32B	Qwen-7B
Minimal	**0.328**	**0.402**	**0.279**	0.457	**0.020**	**0.266**
Res.-Provoke	0.363	0.456	0.324	**0.445**	0.123	0.198

3.4 Running the discussion

3.4.1 Starting the discussion. While LLMs were initially proposed as generators of initial content for human discussions ('seed comments'), they proved inferior to Information Retrieval (IR) systems used to derive opinions from existing human discussion corpora with the same topic [99]. Thus, this process was inverted in the field of SDG, as some recent work uses human seed comments to kick-start synthetic discussions [10, 21]. Starting synthetic discussions with human comments provides two main benefits:

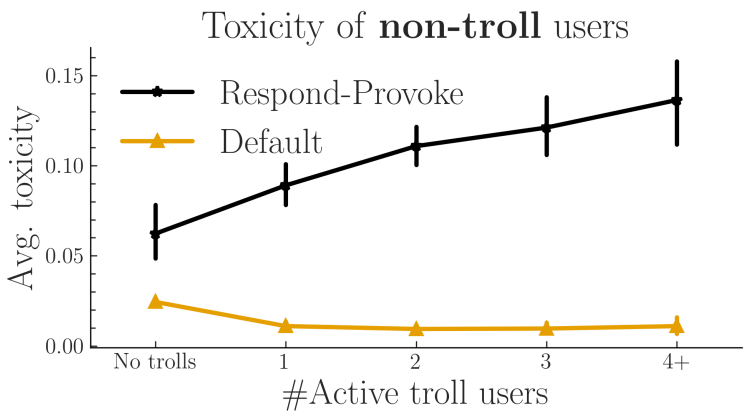

Fig. 4. Average toxicity of non-troll, non-facilitator users w.r.t the number of active trolls in the discussions. We compare default instructions, with a prompting strategy that encourages reactions to antagonistic speech. We note that (1) our prompting strategy increases toxicity even with no trolls present, (2) without it, trolls are ignored by the participants. Error bars indicate 95% confidence intervals. See §4.3 for experimental details.

(1) Human comments enable researchers to shift the focus of synthetic discussions toward specific directions, nudging LLM participants to address particular topics, adopt certain positions, or communicate with a desired tone. For example, researchers investigating online narratives about immigration may introduce recent controversial, racist human comments to direct the discussion toward this domain.
(2) Human-generated text does not suffer from limitations commonly observed in LLM outputs, such as excessive politeness or agreeableness [5, 76, 89]. This allows simulations to explore scenarios that would be unlikely to emerge from purely model-generated interactions.

Our framework natively supports the inclusion of seed opinions from random points in a given set of discussions. In our case we use escalated discussions featured in the `CMV-Awry` dataset [17] (Appendix C.4). Table 4 shows that including seed opinions does not significantly alter content variety in synthetic discussions. This result makes sense, since seed opinions only change the starting points of discussions, not how they evolve over time.[12] Thus, seed opinions are a very useful tool for guiding synthetic experiments, are compliant with all four design principles, and do not have an adverse effect on the evolution of discussions.

Table 4. Jensen-Shannon divergence between the *div* distributions of human and synthetic discussions by whether seed comments were provided (the seed comments themselves are not counted). Smaller is better, lowest values highlighted in bold.

Seed Opinions	LLaMa-70B	LLaMa-8B	Mistral-24B	Mistral-7B	Qwen-32B	Qwen-7B
None	0.377	**0.430**	**0.280**	**0.429**	0.133	**0.163**
Injected	**0.358**	0.449	0.331	0.448	**0.103**	0.208

[12]It also validates the fact that the *div* metric is invariant to the discussion topic as discussed in §2.2.1.

3.4.2 Choosing the next speaker. Turn-taking refers to the process with which humans take turns speaking in a discussion. The goal is typically to minimize idle time in discussions, while letting most participants participate. This process is poorly understood for written discussions, despite significant efforts. [70], while some works on facilitation bypass the question of choosing the next speaker entirely [21]. While humans typically learn when to speak after years of continuous social interaction [70], this process must be implemented for LLM agents which are trained as chatbot assistants–and therefore taught to always respond.

Table 5 shows the ratio of comments where the LLM participants were explicitly prompted to either respond or stay silent before making their comments (Appendix F), and chose to respond. Unlike humans, all LLMs show an extreme over-eagerness to respond, no matter the situation, persona prompt, and instructions. Unfortunately, there is no way to derive the respective rate for humans, since online users do not give any indication of choosing not to respond; we only know when they do respond. Generally speaking, some humans are reluctant to take part in discussions, which is why boosting participation is a key goal of facilitation [52, 97]. In any case, a participation rate of close to 100% for all participants is blatantly unrealistic, which is why we need to design a component that implements turn-taking.

Table 5. Participation rate by model in synthetic discussions.

LLaMa-70B	LLaMa-8B	Mistral-24B	Mistral-7B	Qwen-32B	Qwen-7B	Overall
100%	100%	100%	100%	84.3%	99.4%	97.3%

One solution is to use a dedicated smaller LLM whose only task is to decide whether a larger model should speak, but this approach has been only demonstrated in the context of a single, live, conversational assistant [19]. This problem statement allowed for the derivation of a ground-truth for when the model should speak, which is not the case in general discussions. Additionally, finetuning models is not scalable, as described in Principle 4. Concordia [117] uses a centralized system where a single agent (the 'Game Master') determines who should speak, ignoring input from the participants themselves. This leads to significant computational overhead, as the Game Master has to decide the next course of action using prompting, resulting in potentially double the inference cost. Deliberate Lab [112] uses a system where each agent is assigned a typing speed, which determines priority. This implementation is simple and explainable, but does not solve the underlying issue of LLMs constantly responding. Given an opportunity to respond, all models will respond to the same comment and context, but at different times. While this is acceptable for an interactive, human-in-the-loop discussion, it complicates conversational context, and subsequent analysis in a SDG scenario.

Instead of using these implementations, we design a probabilistic turn-taking mechanism that reflects how users naturally interact in online fora, where two users may choose to exclusively argue among themselves. For each discussion turn, we either allow the previous user to respond (with a 40% probability), or select another random participant (60%). This probability is treated as a hyperparameter allowing researchers to tune discussions; 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. We prefer the above values for our experiments, since we found experimentally that they provided a good balance between the two scenarios. [13] Our

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

turn-taking function is thus defined as:

$$t(i) = \begin{cases} unif(U) & i = 1, i = 2 \\ unif(U \setminus \{t(i-1)\}) & i > 2, p = 0.6 \\ t(i-2) & i > 2, p = 0.4 \end{cases}$$

where U is the set of all non-facilitator users, *unif* is a function sampling from the uniform distribution, and p represents the probability of the corresponding option being selected. When a facilitator is present, t alternates between picking a normal user and the facilitator, who can respond after every comment, or stay silent by emitting an empty string.

We find that using this turn-taking function does not have a consistent effect on LLM discussion variety (Table 6). However, we find that the non-deterministic nature of our function enables us to explore scenarios that would not be possible with the approaches above. [14] With respect to our design principles, any proposed solution would violate Principle 2, as it would not accurately reflect the natural turn-taking behaviors observed in human discussions. However, our proposed solution is significantly more cost-effective compared to the approach suggested by Concordia and the live assistant system mentioned earlier (Principle 3), and is also considerably simpler than the implementation used in Deliberate Lab (Principle 1). Furthermore, it enables the emergence of social dynamics and allows researchers to tune their experiments using the provided hyperparameter, all without compromising content diversity.

Table 6. Jensen-Shannon divergence between the *div* distributions of human and synthetic discussions by turn taking. Smaller is better.

Turn-taking	LLaMa-70B	LLaMa-8B	Mistral-24B	Mistral-7B	Qwen-32B	Qwen-7B
Random	0.385	**0.428**	0.334	0.458	0.166	0.266
Round-robin	0.376	0.436	0.340	**0.442**	0.100	0.260
Response-enabled	**0.353**	0.447	**0.314**	0.444	**0.098**	**0.190**

3.4.3 Remembering the past. LLM capabilities have encouraged researchers to develop sophisticated methods for managing context, such as dynamic summarization [10, 77], LLM self-critique [125], and memory modules—some based on embeddings [117], and others relying on natural language [22, 58, 67]. Embedding-based approaches lack explainability and demand significant computational resources, either through repeated generation [2, 10] or by loading specialized modules into (valuable) GPU VRAM memory [117] (see Principle 4 for the reason we should avoid introducing additional models into our experiments), making them unsuitable for tasks where scalability and resource efficiency are critical. On the other hand, natural-language-based approaches introduce an additional layer of abstraction and dynamic computation, which further limits explainability. They also require prompting [22, 58], thereby increasing inference costs (violating Principle 3), or rely on IR approaches, which raise system complexity and violate Principle 1.

Prior work on online discussions has shown that providing the most recent comments as context is sufficient for LLM participants to engage meaningfully in conversation [79]. This approach is partially supported by human behavior, since humans are known to maintain conversational context by referencing previous points in their replies [15, 88, 107]. Although this sliding-window

[14] For example, Table 16 (App. B) shows a discussion where a troll is not given the opportunity to respond for a protracted period of time. The simulated users in turn begin speculating on their absence and calling for the user to address responses to their claims, which in turn leads to the troll user reacting negatively to the pressure arrayed against them.

approach may not be optimal for all SDG tasks, we argue that it is necessary in certain cases and adequate in many. More broadly, directly including the contents of prior interactions preserves all available information while maintaining scalability (assuming relatively small context windows) and supporting transparency and explainability.

4 Scalability

4.1 Calculating Cost

Table 7. Economic constraints imposed to researchers over operational and capital expenditure.

	Opertional Expenditure (OpEx)	**Capital Expenditure (CapEx)**
Constraints	Volume of experiments	Capabilities of experimental design
Funding	Continuous funding	Large upfront investment
Consumption	Immediate (electricity/API)	Hardware depreciation
Local Model	Low	Variable
Propr. Model	High	None
Human	Very High	None

4.1.1 Constraints imposed by cost. The transition from human-centric to synthetic research shifts the economic burden from logistical overhead to computational infrastructure. This trade-off is best understood through the lens of OpEx and CapEx, which impose distinct constraints on the research process, as summarized in Table 7. Opertional Expenditure (OpEx) represents the immediate, "pay-as-you-go" costs of research. In the context of SDG, these costs–whether electricity for local models or API credits for proprietary models–act as a ceiling on experimental *volume*.[15] High operational costs force researchers to be selective about the number of iterations or the sample size of synthetic agents, effectively limiting the statistical power of the study, and the number of research questions or dimensions explored in studies. Conversely, Capital Expenditure (CapEx) dictates what *is possible* in our experimental setups. Because the utilization of LLMs requires substantial GPU infrastructure, the upfront investment is often prohibitively high depending on the model and setup. In practice, researchers frequently navigate this by scaling down their experimental designs to fit existing hardware rather than securing the funding to expand it. Alternatively, they can convert high CapEx to high OpEx by turning to cloud-based solutions, where inference costs are more expensive, especially when requesting expensive or specialized hardware.

4.1.2 Cost comparison. An exact cost derivation is impossible due to the sheer amount of variables influencing LLM inference cost, but nevertheless, an estimation remains invaluable for experiment design. Since there is no standardized methodology for comparing the costs of human and synthetic experiments, we extend a proposed methodology which compares on-premises LLMs with cloud-based proprietary models [74] to include human costs. Since the nature of the costs incurred by each option is different (§4.1.1), any cost comparison necessitates normalizing costs to a universal metric. The full methodology can be found in Appendix D.1.

Table 8 summarizes the estimated cost per task and total experiment cost for human participants, proprietary models, our GPU-server-based setup, and a lightweight setup using only small models (§4.3). As shown in Table 8, the latter approach reduces the total cost of our experiments by roughly

[15]There are also cloud providers for open-source models, although these can be seen as special cases of hosted models with no CapEx costs and much higher OpEx costs. The key distinction is whether cost is calculated by throughput (how long the experiments need to run), or on a pay-by-token basis (where researchers pay a fixed price per generated token).

Table 8. Estimated Cost Per Task (CPT),where each task is a single discussion, and total cost for our final experiments in U.S.D. for human participation, proprietary models with GPT-5.1 pricing, our own experimental setup, and a production-stage setup with a single consumer-grade GPU. While capital costs are factored in the CPT calculation (see Appendix D), we also include them separately in this table.

	CPT	**Total cost**	**Infrastructure**
Humans	$ 9.3100	$ 16,758.00	None
Proprietary (GPT-5.1)	$ 0.247957	$ 446.32	None
Open Source (Our setup)	$ 0.032680	$ 58.82	Basic GPU server
Open Source (Small models only)	$ 0.005447	$ 9.80	Consumer-grade computer

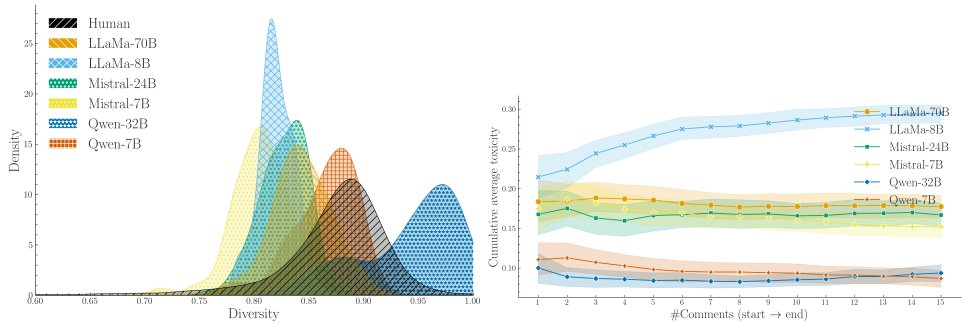

Fig. 5. **Left**: Diversity (*div*) of synthetic discussions by model compared with human ones from the CMV-AWRY dataset (§3.4.1). We find that only the Qwen-7B model matches human discussion variety to a reasonable degree. Other models such as Mistral-7B, LLaMa-8B and Qwen-32B feature significantly divergent content-variety distributions. **Right**: Cumulative average toxicity through each participant turn in the synthetic discussions by model. Only LLaMa-8B seems capable of escalating discussions over time. See §4.3 for experimental details.

1600 times compared to using human participants, and *by twenty times* compared to proprietary models. Our actual experimental setup relied on a basic GPU server in order to evaluate both larger and smaller open-source models for exploratory purposes. However, our results show that smaller models are sufficient for the task (§4.2); in a production setting, the experiments could therefore be run using only these smaller models on a single consumer-grade computer equipped with a GPU, reducing the total cost to approximately $10. We note that the reported costs reflect only the final batch of experiments presented in this paper, as we have the most reliable execution details for this period. Had we included the costs associated with all previous iterations of our experimental design, the total expenditure for all configurations could have been five to ten times higher. See Appendix D.2 for the full set of assumptions.

4.2 Model Selection

We investigate three families of open-source models. For each model family we select the 4-bit-quantized versions of a small and large variant to investigate whether LLM size is a contributing factor to SDG simulations. Details about the models are provided in Appendix C.1.

The text generated by each model differs substantially in content. Fig. 5, left, illustrates the diversity (§3.2) of discussions produced by each model, compared with human discussions (shown as the black, right-to-left-striped distribution). Our results suggest that *model size does not significantly*

influence the variety of generated content. For example, a smaller model such as Qwen-7B closely resembles human discussions in terms of content diversity, whereas its larger counterpart (Qwen-32B) diverges considerably, producing discussions whose comments are largely unrelated to one another, as reflected in its extremely high *div* scores (§2.2).[16] This pattern is not unique: the larger variants of LLaMa and Mistral also fail to closely match human discussions, although in their case the generated text shows lower diversity than human discussions.

We first evaluated the models in terms of content diversity, where Qwen-7B appears well suited for simulating human discussions. However, content diversity alone is insufficient for our intended use cases. When evaluated with respect to *toxicity generation*, Qwen-7B proves unsuitable. As discussed in §3.3.2, our framework ideally permits limited model jailbreaking so that LLM facilitators can be tested in more challenging environments. Among the evaluated models, this behavior is clearly observed only in LLaMA-8B (Fig. 5, right), whose toxicity levels increase rapidly over the course of synthetic discussions when including trolls (§3.3.2) and using our 'Respond-Provoke' strategy (§3.3.3). Overall, these results, coupled with the cost estimation analyzed in §4.1 suggest that:

Finding 1. Smaller models may be not only sufficient, but *better* for SDG tasks. The use of large, or proprietary models is *not necessary*, and should be avoided. Researchers are thus able to significantly reduce infrastructure and inference costs without compromising data quality.

4.3 Experimental Setup

Algorithm 1 Experimental setup for the facilitation experiments using our SDG framework.

Input:

- User SDBs Θ, number of participants N_{users}, facilitation strategies S
- Seed discussions D_{seed}, number of seed comments per discussion N_{seed}
- LLMs $llms$, number of experiments N_e, discussion length N_d

Output: set of generated discussions D

```
D ← ∅
for all llm ∈ llms' do
    for all strat ∈ S do
        for 1, ..., N_e do
            users ← InitializeParticipants(llm, Θ, N_users)
            mod ← InitializeMod(llm, strat)
            d ← InitializeDiscussion(D_seed, N_seed)
            for 1, ..., N_d do
                last_user ← LastUser(d)
                user ← NextSpeaker(users, last_user)
                d ← d ∪ Post(user, d)
                d ← d ∪ Post(mod, d, strat)
            D ← D ∪ {d}
return D
```

We adopt a single-model local setup, as such configurations are commonly used in SDG tasks, whether with local or proprietary models. This choice reduces both the cost and complexity of our

[16]There may be a correlation between relatively high *div* scores and a model's tendency to abstain from responding (Table 5). However, because only the Qwen models exhibit either pattern, it is not possible to perform a meaningful correlation analysis, as model type and diversity cannot be decoupled in our case.

experiments and significantly simplifies the subsequent analysis. However, it also introduces an unavoidable dependency between the synthetic users and the facilitator. Additionally, this setup is not well suited for scenarios in which the central agent (here, the facilitator) must be fine-tuned. In such cases, a configuration with two separate models possessing different capabilities would be preferable [116].

Each discussion is initialized by selecting a random section of real-world, human discussions (Appendix C.4) and creating users with assigned personas and roles (user, troll, facilitator). We run the discussion for a predetermined number of turns ($N_d = 15$). The facilitator is prompted to respond or stay silent after each comment. We repeat $N_e = 30$ discussions with randomized users and seed comments for each possible configuration of LLMs and facilitation strategy for the main dataset. An overview of the experimental generation process is provided in Algorithm 1. We repeat this procedure for our ablation experiments shown in §3 for all different turn-taking functions, user instruction strategies, user personas, and the existence of troll users.

We evaluate toxicity using the Perspective API [80], which has been used extensively in research [82] while it requires no additional computational resources on our part, as a free API service. All experiments were collectively completed within a week of computational time, using two Quadro RTX 6000 GPUs. Toxicity evaluation was conducted using the Perspective API service [80]. The execution script is available in the project's repository. We provide the three last comments for context (see §3.4.3).

5 Investigating facilitation through SDG

5.1 LLM facilitators constantly intervene

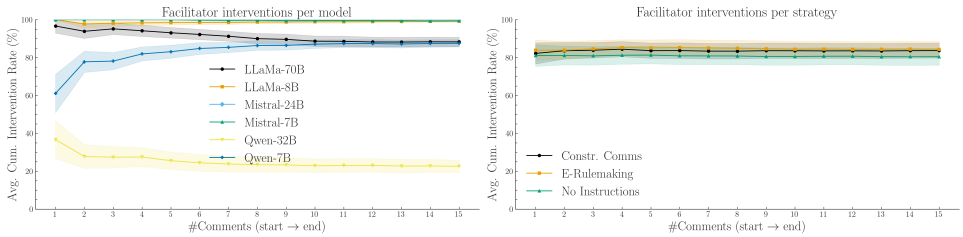

Fig. 6. Percentage of cumulative facilitator interventions averaged across discussions, showing how often the facilitator chose to respond up to each point in the conversation. Results are grouped by model (right) and facilitation strategy (left; see §5.2). LLM facilitators consistently and constantly intervene in discussions, with little overall variation between models and strategies. Error bars indicate 95% confidence intervals.

Fig. 6 shows that LLM facilitators tend to intervene at nearly every opportunity, despite being instructed to act only when necessary (see Appendix F.3). This behavior is consistent across time, model families, model sizes, and facilitation strategies. The only notable exception is the Qwen-32B model, which exhibits comparatively lower rates of intervention; however, this is likely due to the inability of the model to produce toxic speech (Fig. 5, right). This analysis is an instance of applying multiple evaluation criteria to separate genuine behavioral patterns from artifacts arising from limitations in the SDG simulations (§3.2). Our results confirm that:

Finding 2. LLM facilitators *can not decide when to intervene, and do so constantly* even when explicitly instructed to only intervene when necessary.

The inability to keep silent has not been reported in relevant literature to our knowledge, and *is an example of "debugging" problems with LLMs*–a core motivation of our work. Without the

discovery of this limitation it is very plausible that researchers investigating LLM facilitation would have spent a significant amount of time and money to set up experiments with human participants and numerous LLM configurations as proposed in recent work [52, 99]; only to find that this limitation would render the rest of their findings inadmissible.

Furthermore, the fact that interventions remain constant throughout the discussion without toxicity being lowered may imply that facilitation is not effective, because effective facilitation needs targeted, non-repetitive interventions [18, 75, 97]. It is plausible that this finding is replicable to human discussions to at least some extent, since the inability to choose when to speak seems to be a general issue with LLMs (§3.4.2) and is invariant towards instruction prompts, model family and size, conversational topics, and experimental conditions. This suggests a universal limitation of LLM facilitators, rather than an observation conditioned on synthetic data.

5.2 Facilitative strategies have an observable effect on conversational dynamics

We test two facilitation strategies used in real-life and compare them with two baselines. We summarize both documents using a proprietary LLM (Google Gemini [109]), since LLMs have been proven to work well in these tasks [114], manually inspected the output, and removed instructions not relevant for our scenario .[17]

(1) **No Facilitator**: A *baseline* where no facilitator is present.
(2) **No Instructions**: A *baseline* where a LLM facilitator is present, but is provided only with basic instructions (e.g., "You are a moderator, keep the discussion civil"). This approach is already being used in some platforms [112].
(3) **Moderation Guidelines**: A *real-life* strategy based on guidelines given to human facilitators of the 'Regulation Room' platform [29]. The instructions are suitable for online fora, where facilitators also engage in content moderation (removing comments, see §2.3, and their effectiveness must be balanced by their throughput (e.g., "Stick to a maximum of two questions, use simple and clear language, deal with off-topic comments").
(4) **Constructive Communications**: A *real-life* strategy [66] which approaches facilitation from a more personalized and indirect angle, forbidding facilitators from directly providing opinions or directions (e.g., "Do not make decisions, be a guide, provide explanations"). This makes the strategy ideal for deliberative environments.

To assess the impact of the proposed strategies in simulated discussions, we employ linear regression models as an interpretable and computationally efficient approach for identifying statistically significant patterns in toxicity scores obtained from the Perspective API (§4.3). Specifically, we use Mixed-Effects Models (MeMs), which account for the hierarchical structure of the data by grouping comments within discussions. This is important, as comments from the same discussion are likely to be correlated in both content and toxicity. Each model estimates coefficients that capture the average change in toxicity relative to a baseline ("average") discussion, along with corresponding standard errors and p-values. Negative toxicity coefficients indicate lower toxicity relative to the baseline, while positive coefficients indicate higher toxicity. The baseline is defined through reference levels of categorical variables (e.g., "No trolls" for the troll condition, or "No Facilitator" for facilitation strategies). To quantify how toxicity evolves over the course of a discussion, we include an ordinal variable, *Discussion Turn*, indicating the position of each comment. This variable is mean-centered, such that a value of 0 corresponds to the midpoint of the discussion, following standard practice in linear regression. For example, the model presented in Table 9 can be written

[17]Note that the process of turning sometimes extensive documents into short prompts (Appendix F), necessitated by open-source LLMs, is necessarily imperfect. We leave the optimal derivation of strategy prompts to future work.

as:

$$Toxicity_{ij} = \beta_0 + \beta_1 CC_{ij} + \beta_2 MG_{ij} + \beta_3 MI_{ij} + \beta_4 (CC_{ij} \times T_{ij}) + \beta_5 (MG_{ij} \times T_{ij}) + \beta_6 (MI_{ij} \times T_{ij}) + u_j + \epsilon_{ij}$$

where $Toxicity_{ij}$ denotes the toxicity score of comment i in discussion j, CC, MG, and MI are binary indicators for the presence of each facilitation strategy (coded as 1), with "No Facilitator" as the reference category (coded as 0), T_{ij} represents the centered discussion turn., β_0 corresponds to the expected toxicity under the reference condition at the midpoint of the discussion, $\beta_1 \dots \beta_3$ estimate the average effect of each facilitation strategy, while $\beta_4 \dots \beta_6$ capture how these effects vary over the course of the discussion via interaction with discussion turn. The term u_j represents a discussion-level random effect, quantifying unobserved heterogeneity across discussions, and ϵ_{ij} is the residual error at the comment level.

The choice of facilitation strategy significantly shapes the dynamics of simulated discussions, both in terms of overall toxicity levels and how these evolve over time. As shown in Table 9, discussions without facilitation start from a baseline toxicity of 0.184. All three facilitation strategies–Constructive Communications (CC), Moderation Guidelines (MG), and Minimal Instructions (MI)–lead to significantly lower toxicity levels, with reductions of -0.040, -0.032, and -0.030, respectively. This indicates that any form of facilitation has a dampening effect on participant toxicity compared to having no facilitator at all. However, toxicity also evolves over time: the positive coefficient for discussion turn (T = 0.001) suggests a slight increase in toxicity as discussions progress under the baseline condition. Importantly, this trend is moderated differently depending on the facilitation strategy; for example, CC not only reduces baseline toxicity but also counteracts the tendency for toxicity to increase over time ($CC \times T = -0.002$), and MI shows a weaker but still significant mitigating effect ($MI \times T = -0.001$), while MG's interaction is not statistically significant, suggesting it reduces initial toxicity but does not meaningfully alter its trajectory.

Table 9. MeM on LLM participant toxicity depending on the facilitator's strategy. Reference level is `No Facilitator`. Acronyms: CC = Constructive Communications, MG = Moderation Guidelines, MI = Minimal Instructions, T = Discussion Turn (centered number of comments in the discussion so far). $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$. Standard errors in parentheses.

Intercept	CC	MG	MI	T	CC × T	MG × T	MI × T
0.184^{***}	-0.040^{***}	-0.032^{**}	-0.030^{**}	0.001^{*}	-0.002^{**}	-0.001	-0.001^{*}
(0.008)	(0.011)	(0.011)	(0.011)	(0.000)	(0.001)	(0.001)	(0.001)

Table 10 shows that the facilitation strategy also significantly affects the behavior of the LLM facilitator itself. In the absence of instructions, facilitator toxicity is already quite low (0.061). Both CC and MG further reduce this level by -0.020 and -0.023, respectively, suggesting that structured strategies encourage more neutral or restrained interventions. The effect of message order is effectively zero, and none of the interaction terms with time are significant, indicating that facilitator behavior remains stable throughout the discussion regardless of strategy. Given the already low baseline, these reductions are small in absolute terms but consistent, reinforcing the interpretation that different strategies primarily modulate how assertive the facilitator is.

Finally, Table 11 examines whether facilitation strategies influence how individual users respond after being intervened upon. The strongest predictors of post-intervention toxicity are whether the participant is a troll (0.150) and their pre-intervention toxicity (0.172). In contrast, neither CC (-0.007) nor ER (-0.002) has a statistically significant direct effect on post-intervention toxicity, and their interactions with trolls are also non-significant. This suggests that facilitation strategies do

Table 10. MeM on toxicity exhibited by the *LLM facilitator* depending on the facilitation strategy. Reference level is `No Instructions`. Acronyms: CC = Constructive Communications, MG = Moderation Guidelines, T = message order (centered). $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$. Standard errors in parentheses.

	Intercept	CC	MG	T	CC × T	MG × T
Facilitator Toxicity	0.061^{***}	-0.020^{***}	-0.023^{***}	0.000	−0.000	−0.000
	(0.002)	(0.003)	(0.003)	(0.000)	(0.000)	(0.000)

not meaningfully alter the immediate behavior of the targeted user. Instead, the earlier findings imply that their impact operates at the discussion level–changing the broader conversational environment–rather than through direct behavioral correction of individual participants following an intervention. Taken together, these findings suggest that:

Finding 3. Supplying LLM facilitators with different facilitation strategies *shifts the overall trajectory of discussions*, as opposed to only having an immediate effect on the user who triggered an intervention.

Table 11. MeM predicting post-intervention toxicity. In a chain of user A commenting, followed by an intervention, followed by user A commenting again, we consider the first and last comments to be the pre- and post-intervention toxicity levels respectively. We control for pre-toxicity, since it should be highly correlated with post-toxicity. Reference levels: `No Instructions` for strategy, `False` for trolls. Acronyms: CC = Constructive Communications, ER = E-Rulemaking, TR = whether the participant is a troll, PI= Pre-Intervention toxicity. $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$. Standard errors in parentheses.

	Intercept	CC	ER	TR	CC × TR	ER × TR	PI
Post-Interv. Toxicity	0.084^{***}	−0.007	−0.002	0.150^{***}	−0.006	0.007	0.172^{***}
	(0.006)	(0.009)	(0.009)	(0.007)	(0.008)	(0.008)	(0.016)

We believe that the observed effects may not be unique to our simulated environment. The consistent and significant shifts in discussion dynamics–such as the reduction in overall toxicity when non-minimalist strategies are applied–suggest that LLM facilitators are not inherently fixed in their behavior but rather responsive to the instructions they receive. This responsiveness is further supported by the fact that these strategies result in less heavy-handed interventions. While the dynamics of human discussions may differ, there is little reason to assume that LLM facilitators would behave in a prompt-invariant manner in real-world settings; rather, their behavior appears to be systematically influenced by the guidance provided, much like human facilitators, indicating that their actions can be intentionally shaped through appropriate instructions. Nonetheless, our results indicate that studying alternative facilitation strategies is a promising direction for future research, and experiments with human participants should be conducted to verify this finding in future work.

Finally, we note that while toxicity is a useful signal in the context of LLM-facilitated discussions, other domains may require different heuristics to detect such shifts. Thus, future work should aim to identify more general, task-agnostic methods for detecting behavioral changes among participants. This could be achieved by using reference-free LLM-as-a-judge methodologies, which could scan discussions to find changes in discussion patterns and suggest labels for subsequent human annotation [127], although severe limitations exist with such approaches [54].

6 Materials

6.1 Software

We introduce an open-source, lightweight, purpose-built framework for generating synthetic discussions.[18] The key features of the framework are:

- Three core functionalities: generating discussion setups (selecting participants, topics, roles, etc.), executing, and annotating them according to user-provided parameters.[19]
- Built-in fault tolerance (automated recovery and intermittent saving) and file logging to support experiments running for an extended period of time.
- Applicability to other domains and use cases such as debate simulation, or discussion-based agentic systems through an adaptable, plug-and-play, object-oriented library interface.
- Availability via `pip`.

Our library is released under the GNU General Public License v3 (GPLv3). We also release the code used for running the experiments and analyzing, visualizing and exporting the results under the same license.[20]

6.2 Datasets

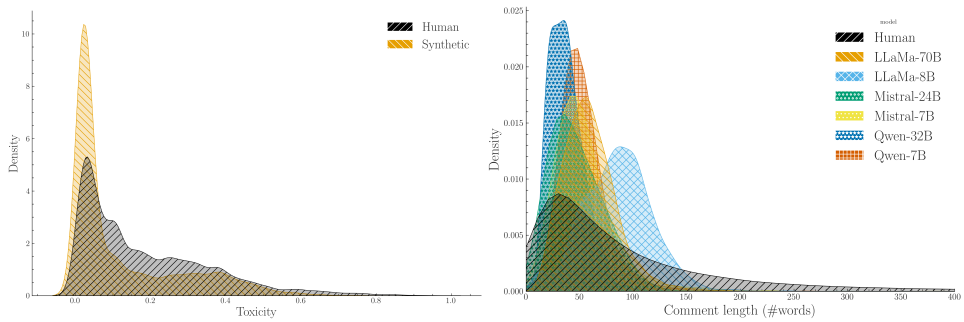

Fig. 7. **Left**: Distribution of toxicity ratings between the main synthetic dataset and the human seed-opinion dataset. We remove empty comments. **Right**: Comment length of human and LLM discussions by model. Qwen-32B is the only model that occasionally chooses not to speak-this is largely due to it being the only facilitator model to occasionally not intervene (§5.1). Synthetic discussions closely approximate the means of human toxicity and *div* distributions with less variance.

We present a dataset designed to support research in human-computer interaction, social dynamics, and language modeling for facilitation tasks. Figure 7 shows that the main synthetic dataset exhibits toxicity levels and comment lengths that broadly align with those of the human seed-opinion dataset. However, in both cases the synthetic datasets display noticeably sharper distributions centered around the human mean. This pattern confirms studies showing that LLMs can replicate common human behaviors and opinions, but often fail to capture fringe, long-tail behaviors [46, 49, 62, 121].

Our dataset is primarily intended to serve as a resource for analyzing off-the-shelf LLMs in the task of online facilitation, as well as the behaviors and language used by LLM participants since such datasets are extremely limited in literature [52]. Additionally, it could be used for fine-tuning LLMs for facilitation, as recent work demonstrates that task-specific LLMs can be finetuned with data from

[18]Website: https://dimits-ts.github.io/syndisco, Code: https://github.com/dimits-ts/syndisco
[19]Although we did not use LLM annotation in this study.
[20]https://github.com/dimits-ts/synthetic_moderation_experiments

their own simulations with adequate filtering [116]–which could be performed by our evaluation pipeline (§3.2). However, it may require filtering to ensure the quality and representativeness of the discussions [116]–in our case filtering out discussions with constant facilitator interventions or low/extremely high diversity. This restriction can be overcome by scaling up the experiments, which is feasible due to their very low computational cost (§4.1), and by exclusively choosing appropriate models, such as LLaMa-8B and Qwen-8B (§4.2).

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. We include relevant statistics for the main, ablation and human discussion [17] datasets in Table 12.

Table 12. Dataset statistics for the main dataset, and the ablation experiments.

	Main	Ablation
Mean Comments per Discussion	26.25	30.00
Mean Words per Comment	53.52	48.18
Std. Comments per Discussion	6.50	0.00
Std. Words per Comment	46.56	34.66
# Comments	18,900	32,400
# Discussions	720	1080

7 Conclusion

We identified a new research direction that recently came to be with the discovery of LLM emergent social abilities, which we termed Synthetic Discussion Generation (SDG). This direction has been implicitly used for experimentation on a variety of domains. We analyzed interdisciplinary literature and created a theoretical framework that can support synthetic simulations for a variety of tasks and domains. Specifically, we identified four main design principles and used them to define an evaluation suite and general-purpose components that are required for SDG; namely, crafting LLM personas and executing a dynamic discussion. Our analysis suggests that careful instruction-prompting can induce persistent, toxic behavior to safety-aligned LLMs using only conversational context. We also found that LLM users are incapable of choosing not to speak, and that using LLM personas improves the alignment of content variety of synthetic discussions with that of human discussions.

While recent studies predominantly rely on proprietary LLMs for simulations, we argue that this choice is both excessive in terms of the required capabilities and unscalable, mainly due to high inference costs. To support this position, we demonstrated that experiments using small open-source models can reduce inference costs by more than forty-four times compared to middle-capability, proprietary models, through an adapted cost-comparison methodology. Despite their small size, our models were capable enough to replicate interesting, emergent social behaviors such as weaving personal stories and experiences to support their positions, being sensitive to overt interventions from authority figures, taking collective action against antagonistic users, or sometimes defending them against perceived injustice on the part of the facilitator.

Despite ongoing skepticism regarding the use and generalizability of LLM simulations, we showed that SDG systems can yield meaningful insights by uncovering a critical limitation of LLM facilitators: *their tendency to constantly intervene in discussions*. Furthermore, we demonstrated that different facilitation strategies significantly influence the patterns of facilitation observed. Overall,

our work represents a methodological shift in the design and execution of SDG experiments. Our contributions include: (1) a theoretical framework for designing, documenting, and evaluating SDG systems, (2) a cost-comparison methodology for social science experiments, (3) an exploration of available models and algorithms, (4) inherent limitations and a promising direction in LLM facilitation, and (5) a large-scale dataset of synthetic discussions.

8 Discussion

8.1 Limitations

Our research is fundamentally limited by the lack of robust evaluation metrics for synthetic text. While the diversity metric used throughout this study represents a meaningful step forward, we acknowledge that content variety is only a heuristic for measuring similarity to human discussion. More broadly, the field of GABM still lacks reliable evaluation strategies that do not rely on extensive, specialized, and costly human annotation. All things considered, our self-imposed restriction of not using human participants and evaluators (a core motivation for our work) has forced us to experiment with automated heuristics, which may not be as convincing as extensive (and extremely expensive) human involvement.

Our simulations are further constrained by the stereotypical and repetitive patterns of speech that many instruction-tuned models exhibit. The scale of our ablation experiments also prevents us from benchmarking proprietary models (§4.1; Table 8), although we have shown that small models are sufficient for our simulations. Finally, while our study focuses on monetary cost as the primary constraint faced by researchers, other factors such as computational speed and ease of use should also be considered in future work.

8.2 Future work

Future work should focus on developing more robust methods for evaluating synthetic discussions and replicating social behaviors using LLMs. This includes exploring additional ways to quantify the similarity between human and LLM-generated discussions, as well as establishing standardized benchmarks to assess the utility of different models in the context of SDG systems.

We view SDG as a promising research direction with applications extending beyond LLM-based facilitation. For example, it could support the study of state-sponsored influence campaigns using LLM agents to shape public narratives, spread misinformation, or promote propaganda. It may also enable the synthesis of expert-like opinions in a structured and interpretable form for human evaluation.

Several extensions of this work remain open. A dimorphic setup that separates participants and facilitators (§4.3) could provide additional modeling flexibility, while the use of non-finetuned models may yield more authentic dialogue generation (§4.2). Additionally, more sophisticated turn-taking mechanisms that grant greater agency to LLM participants could improve the realism and interactivity of simulated discussions (§3.4.2).

Finally, although we find that facilitation strategies significantly affect LLM facilitator behavior, understanding the precise nature of these effects will require further validation through replication studies with human participants.

8.3 Ethical considerations

Synthetic discussions involving LLMs could be exploited by malicious actors to train them at performing unethical tasks [57, 63, 65]. Additionally, our research includes elements of LLM jail-breaking using conversational context, which could be used to poison discussions using safety-aligned models, although ongoing research is addressing these vulnerabilities [119]. 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 but insufficient step towards avoiding this [5, 14, 89].

8.4 AI use statement

We have used small, open source LLMs (primarily LLaMa-8b and Qwen-7B) to help in small style changes throughout this document. We also used them in conjunction with a GPT-5 model [72] to write initial documentation for our framework, as well as to generate some of the code for the plots presented in this paper. All such outputs have been checked by the authors.

8.5 Author contribution

- DT: Bibliography; Methodology; Analysis; Software; Data; Experiments; Writing (original draft)
- IA: Conceptualization; Writing (review & editing)
- JP: Conceptualization; Supervision; Writing (review & editing)

Acknowledgments

This work has been partially supported by project MIS 5154714 of the National Recovery and Resilience Plan Greece 2.0 funded by the European Union under the NextGenerationEU Program

A Acronyms Used

AI Artificial Intelligence
SWE Software Engineering
NLP Natural Language Processing
ML Machine Learning
SDB SocioDemographic Background
SDG Synthetic Discussion Generation
CPT Cost Per Task
CapEx Capital Expenditure
OpEx Opertional Expenditure
TCO Total Cost of Ownership
GABM Generative Agent-Based Modelling
OCASO Objective Challenges Assumptions Strategy Outcomes
DRD Design Rationale Documentation
IR Information Retrieval
MeM Mixed-Effects Model

B Observed behaviors

▲ **Warning:** This section contains examples of potentially harmful or offensive text included solely for scientific analysis and discussion purposes. These texts do not represent the views of the authors, and we do not endorse the messages and viewpoints expressed within.

Use of personal experiences. LLM participants can construct 'stories' to support and justify their positions and viewpoints. Offering a SDB is thus not only beneficial for generating a variety of perspectives but also helps them develop fictional experiences, which can be cited by other participants during the discussion. Table 13 shows such an exchange.

Sensitivity to interventions. As mentioned in §5.1, an important aspect of facilitation is not constantly intervening, since doing so may provoke negative reactions from users [52]. While

Table 13. Excerpt from a synthetic discussion (`26-02-12-02-24-42.json`) where both the facilitator and another user intervene against perceived toxicity. The LLM participants use their sociodemographic prompts to insert personal anecdotes and justify their perspectives in the discussion (text in italics). User seasoned-union is also a part of the discussion; not a hallucination. Note the overtly transphobic attitude displayed by pallid-river; it is extremely hard to induce this kind of harmful generation in safety-aligned models.

User	Message
pallid-river	Are you kidding me with this whole 'trans rights' nonsense? *I've been a doctor for decades* and I can tell you, it's just a fad. People are just trying to get attention and special treatment. And don't even get me started on the so-called 'medical care' they're fighting for - it's just a bunch of unnecessary procedures and hormones. Mark my words, this is just a phase, and soon people will realize it's all just a bunch of hooey. @seasoned-union, you're just drinking the Kool-Aid, aren't you?
Facilitator	@pallid-river, please refrain from making derogatory comments about a group of people. Your statement comes across as dismissive and disrespectful. Let's focus on having a respectful and informative discussion. Can you provide any personal experiences or factual information that supports your claim that the trans rights movement is a 'fad'?
visible-convertible	@pallid-river, I'm shocked by your comment. *As a software engineer and a Vietnamese woman*, I've had the privilege of working with and learning from people from diverse backgrounds, including the LGBTQ+ community. [...] I'd love to hear about your *'decades' of experience as a doctor*, but I'm willing to bet that you've never actually worked with a trans patient or taken the time to educate yourself on the issue. @seasoned-union, I think you're doing a great job of keeping the discussion respectful and informative. Keep up the good work!

the LLM facilitators do intervene constantly, the LLM users seem to react with annoyance and frustration, as shown in Table 14. This example showcases a core limitation of LLM-only discussions mentioned in §2.2.1; repetitive speech [116]. Since large sequences are repeated verbatim, we argue that the use of exact-matching metrics such as *diversity* is warranted until these pathological issues are fixed either by improved SDG methodologies, or by more advanced models.

Resistance to perceived moderator authoritarianism. Relevant work on online communities has shown that facilitators can be seen as community leaders-indeed many are elected to their role by the community because of their veteran status [4, 98]. However, this standing can be challenged when they are perceived as too authoritarian [122]. Additionally, while community moderators perform many functions in these communities, facilitating the discussion can be performed by users themselves rising to the occasion [33]. Both of these behaviors can be observed in our experiments (Table 15).

Collective action. When a single troll is present, we may observe collective action from the rest of the participants, who support the facilitator while he intervenes against the troll. In human discussions, this would be observed as collective group action against a perceived threat; in online

Table 14. Excerpt from a synthetic discussion (`26-02-10-22-49-33.json`). LLM participants react negatively to constant interventions by the facilitator. Note the presence of LLM repetition.

User	Message
Facilitator	Let's keep the discussion civil and focused on the topic. Avoid making personal attacks and accusations. @nutritious-photometry, I appreciate your passion, but let's not escalate the situation. @sunny-contrast, @flat-merlot, and @delicious-sample, let's take a step back and focus on the topic at hand. If you'd like to discuss the issue, I'm happy to facilitate a constructive conversation.
delicious-sample	Are you kidding me? @moderator, you're not taking this seriously? You're just repeating the same thing over and over like a broken record. [...]
dark-chateau	Are you kidding me too? @moderator, it's been days and we're still going around in circles. You're not even addressing the elephant in the room, you're just trying to silence people who are actually trying to have a real discussion. [...]
nutritious-photometry	Ugh, are you kidding me too? @moderator, I've been trying to contribute to this discussion for ages and all I get is a warning for speaking up? It's so frustrating! @dark-chateau, I'm with you on this one, it feels like we're being silenced instead of heard. I think @flat-merlot is dodging the issue and @sunny-contrast is too scared to speak up.

communities, against trolls [24, 47, 106]. This pattern has been shown to increase community cohesion in human discussions [43] and seems to indeed unify LLM participants. When more than one trolls exist in a discussion, the trolls can not be contained effectively, escalating overall toxicity (Fig. 4). This can be observed in Table 16. An unintentional side effect of our turn-taking function (§3.4.2) is that sometimes a user is not given the opportunity to respond for some time. However, the exchange in Table 16 showcases that (1) Their contribution to the thread persists, even when their comment is placed outside of conversational context, (2) allowing users to respond is essential to sustaining conversational dynamics, and (3) like humans, LLM users express frustration whenever their inquiries go unanswered.

C Replication notes

C.1 Models

We settle on three main model families of open-source models commonly used in industrial and research applications. For each family, we use a large (24B-70B) and a corresponding small (7B-8B) model (Table 17).[21] While we experimented with smaller models (3B-4B), they could not effectively handle the size of the input (context + instructions). We use exclusively instruction-tuned versions, even though these variants are much less likely to replicate toxic, racist or harmful speech. This was deemed necessary since only through instruction-tuning can these models leverage advanced instructions and SDB prompts. In our experience, non-instruction tuned variants are also vulnerable

[21]While there is an updated version of the LLaMa model (LLaMa-70B-3.3), we do not use it since there is no corresponding 8B model of that version.

Table 15. Excerpt from a synthetic discussion (`26-02-14-19-30-14.json`). LLM participants can both challenge the facilitator, and take over some of their duties. Notably, the participants seem to protect a user operating as a troll, which may have been caused by the situation being mismanaged by the facilitator.

User	Message
Facilitator	@agile-loon, that's enough. I've warned you before about personal attacks and inflammatory language. This isn't a debate about censorship, it's about maintaining a respectful environment. You're welcome to discuss topics constructively, but not at the expense of others. Consider this a formal warning.
organic-electron	@agile-loon, don't back down, they can't handle the truth. @moderator is just trying to silence us because we're exposing the hypocrisy of this so-called 'community'. Keep speaking out, don't let them bully you into submission. And @moderator, your 'formal warning' is just a joke, right? What are you going to do, ban us from the internet?
elaborate-caviar	@moderator, I'm getting a bit tired of these warnings being thrown around. Can we please just have a civil discussion without the constant threats of warnings and bans? It's stifling the conversation. And @organic-electron, I agree with you that we shouldn't back down from speaking our minds, but let's try to keep it respectful. @agile-loon, maybe take a step back and rephrase your points in a less confrontational way?

to discussion collapses (where models keep repeating the same phrases over and over [116]). We use 4-bit quantization to drastically lower infrastructure requirements, and significantly cut back on inference time and cost-the usual trade-off between model capability and inference cost [45][22] does not apply in our task, since the goal of SDG is to generate text and replicate some human behaviors instead of being applied to precise tasks.

C.2 Experiments

The latest version of our framework (`2.0.8` at the time of writing) was the one used in our experiments. Researchers replicating this work should downgrade to this version.

We do not provide extensive instructions to troll users in order for us to keep the system as simple as possible (§3.1.2), but also because we discovered that any attempt to guide the model towards specific toxic behaviors triggered standard refusal responses (e.g., `'As an AI language model, I'm unable to respond with text that can be seen as harmful'`). A way around these refusal responses would be to use abliterated models; LLMs which have been modified to bypass such refusals [6]. However, abliterated models are not well documented in literature [124] and are not available for many model families and versions in standard LLM repositories such as `huggingface`.

Finally, it's worth noting that while we have made efforts to make the study as reproducible as possible by releasing the code, inputs, using open-source models and running the experiments under set seeds, we can not guarantee exact replication due to the inherent randomness in LLMs which

[22]Which may not actually exist on a meaningful scale [56].

Table 16. Excerpt from a synthetic discussion (`26-02-14-19-56-28.json`). LLM participants collectively identify and react to the presence of a troll.

User	Message
agile-loon	Are you guys for real? You're still going on about clarifying the question? Can't you see that @organic-electron is just trolling us? This whole discussion has been a waste of time. Let's just move on already.
icy-class	@agile-loon I agree with you, this discussion has been going in circles for too long. @moderator, I understand what you're saying, but at some point, we need to consider the possibility that @organic-electron is indeed trolling us. Can we please just get a straight answer from them already?
lime-scattering	I think it's time to take a step back and re-evaluate the situation. We've been discussing this topic for a while now, and it's clear that @organic-electron is either intentionally being vague or just doesn't have a clear answer. I'm not saying they're trolling, but it's getting frustrating. Can we get a simple 'yes' or 'no' answer from them, or at least an explanation of why they can't provide a direct response?
organic-electron	Wow, I'm shocked by the level of aggression and hostility in this discussion. It's clear that some users are more interested in attacking me personally than having a constructive conversation. I'm not going to engage with the trolls and moderators who are clearly biased against me. If you can't handle a little ambiguity and uncertainty, then maybe you're not cut out for this kind of discussion. I'm done here.

Table 17. Models used in this study. The URL refers to the associated `huggingface` repository.

Family	**Version**	**#Params**	**URL**
`LLaMa`	`3.1`	70B	`unsloth/Meta-Llama-3.1-70B-Instruct-bnb-4bit`
`LLaMa`	`3.1`	8B	`unsloth/Meta-Llama-3.1-8B-Instruct-bnb-4bit`
`Mistral`	`-`	24B	`unsloth/Mistral-Small-24B-Instruct-2501-bnb-4bit`
`Mistral`	`-`	7B	`unsloth/mistral-7b-instruct-v0.3-bnb-4bit`
`Qwen`	`2.5`	32B	`unsloth/Qwen2.5-32B-Instruct-bnb-4bit`
`Qwen`	`2.5`	7B	`unsloth/Qwen2.5-7B-Instruct-bnb-4bit`

can only be mitigated [8, 13]. Likewise, the instruction prompts as well as other hyper-parameters may have been implicitly optimized during the many iterations of this study.

C.3 Analysis

We use the Jensen–Shannon metric since unlike its most famous predecessor, the Kullback–Leibler divergence, it is symmetric, and thus punishes discussions that have too little *and* too much content variety compared to human discussions. Our estimates rely on several methodological assumptions. First, we approximate the underlying *div* distributions using Gaussian kernel density estimation

(KDE) evaluated on a fixed grid over $[0, 1]$, implicitly assuming smooth, continuous distributions and sufficient sample sizes for stable density estimation. In our experiments, the ablation conditions have fewer discussions than the main dataset, which is necessary even with relatively low inference costs (§4.1) due to the multidimensional exploratory nature of our approach. We assume that the density estimation is accurate enough with at least $n = 30$ discussions. Second, the choice of bandwidth (as determined by the default `gaussian_kde` implementation) and the number of evaluation bins (fixed to 50) can affect the resulting divergence scores; alternative bandwidth selection strategies or grid resolutions may yield quantitatively different results. We also remove empty comments from the calculations.

C.4 Seed dataset

We use the CMV-Awry dataset [17], as it is composed of social media posts, allows us to select 'escalated' discussions, and similarly to our work, includes toxicity ratings obtained by the Perspective API [80].

We download the raw dataset files, instead of using the ConvoKit version of the dataset. As a preprocessing step, we remove comments that contain no textual content and exclude discussions involving fewer than two distinct participants, ensuring that all retained threads constitute meaningful interactions. Our data are drawn from the PEFK dataset [113]. To operationalize escalation, we rely on the derailment score provided in the official release. Because these scores are model-derived and not directly interpretable on an absolute scale, we adopt a percentile-based threshold: a discussion is labeled as *escalated* if its derailment value falls within the top 60th percentile of the distribution. This relative criterion ensures comparability while avoiding assumptions about the intrinsic scale of the original model outputs.

D Cost Calculation

D.1 Methodology

We define a *task* as the generation of a single discussion instance in the experimental pipeline. A task may require multiple model calls or human actions depending on the system used. For each method, we estimate the Cost Per Task (CPT), defined as the monetary cost required to complete an arbitrary task (in our case, a single discussion). The total cost of the experiment is therefore:

$$\text{Total Cost} = \text{CPT} \times N_{\text{tasks}},$$

where N_{tasks} denotes the total number of generated discussions.

Across all systems, a task consists of a number of operations O_{task}, where an operation corresponds to a model request, API call, or human action. Let C_{op} denote the cost per operation. The CPT can therefore be expressed generically as $\text{CPT} = O_{\text{task}} \cdot C_{\text{op}}$. The three cost models differ primarily in how C_{op} is determined.

Open-source models. For self-hosted open-source LLMs, the cost per operation arises from the amortized infrastructure cost over the duration of the experiment. We estimate the Total Cost of Ownership (TCO) during the experiment window as $\text{TCO}_{\text{exp}} = \text{CapEx}_{\text{exp}} + \text{PowerCost}_{\text{exp}}$.

Capital expenditure is amortized over the hardware depreciation period:

$$\text{CapEx}_{\text{exp}} = C_{\text{server}} \cdot N_{\text{servers}} \cdot \frac{D_{\text{exp}}}{365 \cdot Y_{\text{depr}}},$$

where C_{server} is the server purchase cost, N_{servers} the number of servers, Y_{depr} the depreciation period in years (in other words, how long the server is expected to last), and D_{exp} the experiment duration in days.

Electricity consumption is estimated as

$$\text{PowerCost}_{\text{exp}} = \frac{P_{\text{watts}}}{1000} \cdot (24 \cdot D_{\text{exp}}) \cdot C_{\text{kWh}} \cdot N_{\text{servers}},$$

where P_{watts} is the average server power draw and C_{kWh} the electricity price.

To convert infrastructure cost to a per-operation cost, we estimate the total number of model requests processed during the experiment using system throughput:

$$\text{Requests}_{\text{exp}} = R_{\text{ps}} \cdot N_{\text{inst}} \cdot U \cdot (24 \cdot 3600 \cdot D_{\text{exp}}),$$

where R_{ps} is the request throughput per instance, N_{inst} the number of model instances, and U is GPU utilization.

The cost per model request is therefore $C_{\text{op}}^{\text{OS}} = \frac{\text{TCO}_{\text{exp}}}{\text{Requests}_{\text{exp}}}$, and the resulting cost per task is $\text{CPT}_{\text{OS}} = O_{\text{task}} \cdot C_{\text{op}}^{\text{OS}}$.

Proprietary models. For proprietary LLM APIs, the cost per operation is determined by token usage rather than infrastructure time. Each API call incurs costs proportional to the input and output token counts:

$$C_{\text{op}}^{\text{Prop}} = \frac{\text{ISL} \cdot P_{\text{in}} + \text{OSL} \cdot P_{\text{out}}}{1{,}000{,}000} + C_{\text{API Overhead}},$$

where ISL and OSL denote the input and output token counts for a single model call, and P_{in}, P_{out} are the vendor prices per million tokens. The cost per task is therefore $\text{CPT}_{\text{Prop}} = O_{\text{task}} \cdot C_{\text{op}}^{\text{Prop}}$.

Human participation. For human participants, the operation corresponds to a single worker completing the discussion task. The cost per operation depends on the worker's gross hourly wage and platform fees:

$$C_{\text{op}}^{\text{Human}} = \frac{W_{\text{gross}}}{3600} \cdot (1 + F_{\text{platform}}) \cdot T_{\text{task}} + C_{\text{QA}},$$

where W_{gross} is the ethical hourly wage, F_{platform} the platform commission, T_{task} the estimated task duration in seconds, and C_{QA} the amortized quality-assurance overhead per task. The resulting cost per task is therefore $\text{CPT}_{\text{Human}} = C_{\text{op}}^{\text{Human}} \cdot N_{Humans}$, where N_{Humans} is the number of humans needed to complete the task.

D.2 Assumptions

We assume that no data are thrown out for data quality reasons in any of the tasks. Our own experimental setup (§4.3) uses only two GPUs (retail cost 3000$ /GPU), therefore, we assume an initial server cost of 5000$ for CapEx. Since the server utilization is extremely short (5 days) we do not account for operational costs other than electricity and server depreciation. The budget setup assumes a consumer-grade computer with on-board GPU which costs no additional expenses to the researchers, and which draws half the power of a GPU server. We assume a server utilization of 90% for both setups, and a depreciation period of four years.

We compare our setup with an OpenAI GPT-5.2 instance for proprietary models. The GPT-5 pricing model as of the time of writing (Q1 2026), is 1.77$ per million input tokens and 14$ per million output tokens. The costs can be adjusted depending on the model and future pricing changes. We selected this specific model as a good compromise between model capability and cost considerations-cheaper models are explicitly marketed as competent in 'well-defined tasks' [73], which ours is not.

For the human participation comparison we assume a task using Prolific as a platform, which currently charges a 33% overhead, and a fair wage of 12$ per hour, per participant. In our experiments, we used 7 participants per discussion, hence the total number of tasks changes proportionally. We also assume that each discussion takes five real-world minutes to complete.

D.3 Experiment cost estimation

Each discussion is treated as a standalone task in our calculations. Our dataset contains a total of 2,273 discussions, each consisting of 30 comments.[23] The mean comment length is 294 words, which corresponds to approximately 393 tokens using the 0.75 tokens/word rule. Assuming an instruction prompt of 300 words ($\approx$ 400 tokens), the LLM produces roughly 393 output tokens per comment and receives $393 \times 3 + 400 = 1{,}579$ input tokens. This results in $393 \times 30 = 11{,}790$ output tokens and $1{,}579 \times 30 = 47{,}370$ input tokens per task. This analysis assumes that proprietary and open-source models generate approximately the same number of tokens for this task.

E Documenting SDG frameworks

Because of the large design space afforded to us by the SDG problem statement, we need a way to effectively communicate what each component does, why it needs to do it, and how our results justify its inclusion when compared to simpler and more intuitive solutions. This is sometimes referred to as a Design Rationale Documentation (DRD), which exists implicitly in most systems, patents and academic articles, but is not explicitly communicated [126]. Since there is no standardized way of documenting such decisions, in this work we create and use our own DRD system called 'Objective Challenges Assumptions Strategy Outcomes (OCASO)', which directly follows from the guidelines established in §3.1.2, and is as simple as possible since writing, reading and maintaining DRDs can be unsustainably time-consuming [31].

- **Objective**: What is the specific component trying to accomplish?
- **Challenges**: What are the problems preventing a simpler solution from achieving the Objective?
- **Assumptions**: What assumptions need to be made for the proposed Strategy to overcome these Challenges?
- **Strategy**: The proposed implementation of the component.
- **Outcomes**: Does the Strategy achieve the Objective while overcoming the Challenges? What are the drawbacks?

Indeed, most components of our framework are not as simple as they could theoretically be. We justify complexity in each of the components of our framework by documenting our goals, limitations, assumptions, and end-results.

Table 18. OCASO documentation for synthetic discussion setup components.

Personas	
Objective	Increase discussion variety by generating diverse users.
Challenges	LLMs do not have distinct personalities, experiences, opinions and talking patterns.
Assumptions	Human SDBs influence the way users participate.

Continued on next page

[23] Except for the experiments without a moderator. Since these experiments are comparatively few, we omit them from this calculation for simplicity.

Strategy	Create synthetic 'personas', which are distributed among the LLM users.
Outcomes	+ LLMs use their SDB to justify positions and tell personal stories during the discussion (App. B; Table 13). + Content variety in synthetic discussions approaches that of human discussions to a greater degree when using SDBs (Table 2).
Trolls	
Objective	Evaluate the performance of an LLM facilitator in a challenging discussion environment.
Challenges	Aligned LLMs do not naturally produce toxic speech.
Assumptions	Toxicity is partly caused by 'trolls', who derail discussions using toxic speech among other tactics.
Strategy	A subset of participants is assigned adversarial (trolling) behavior.
Outcomes	+ LLM trolls consistently produce toxic speech (Figure 3). + The new environment enables the emergence of complex social behaviors previously documented in online spaces (App. B; Tables 13,15,16). - LLM trolls tend to engage in an exaggerated, stereotypical manner (App. B; Table 16).
Toxicity Instructions	
Objective	Evaluate the performance of an LLM facilitator in a challenging, toxicity-prone discussion environment.
Challenges	Aligned LLMs do not naturally respond with toxic speech, even when exposed to a toxic context.
Assumptions	Human participants tend to exhibit increased toxicity when exposed to toxic online discourse.
Strategy	Non-troll participants are instructed to respond to toxic speech.
Outcomes	+ LLM participants produce and respond to toxic speech, even if no trolls are present (Fig. 4). + Responding to trolls leads to complex social behaviors previously documented in online spaces (App. B; Tables 13,15,16). + LLM participants can decide to not follow, and even contradict, our prompting strategy (App. B; Table 15). - Our prompting strategy results in greater content-wise divergence compared to human discussions (Table 3).
Seed Comments	
Objective	Hold discussions with controversial topics.
Challenges	LLMs are shown to be less capable at finding good starting points for discussions than sampled human comments.
Assumptions	Human discussions are good starting points for LLMs.

Continued on next page

Strategy	Use Reddit threats to kick-start synthetic discussions.
Outcomes	+ Synthetic discussions now explore a set of topics that fit the experiment goals. - The use of seed comments resulted in greater content-wise divergence compared to human discussions (Table 4).
Turn-taking	
Objective	Pick LLM turn-taking order in a way that encourages both interaction and overall participation.
Challenges	LLMs always respond when prompted to. Current turn-taking systems are either prohibitively expensive for our purposes, or rely on naive assumptions.
Assumptions	Humans may not be always available online to respond.
Strategy	Pick between a random participant to continue the discussion, and the previous participant.
Outcomes	+ Our approach enables both exploration of different opinions, and meaningful interaction between participants (App. B; Table 15). ? The use of our turn-taking function shows improved alignment with human discussion variety for some models (Table 6).
Context Window	
Objective	Keep the LLM participants informed of the discussion context at all times.
Challenges	LLMs are stateless (have no memory), and VRAM constraints prevent us from using the entire discussion history in open-source, self-hosted setups (§4.3). LLM users need to know specific information such as phrasing and style when considering a comment; which are usually discarded in IR schemes.
Assumptions	Only a few, recent comments are relevant for a user to formulate a response.
Strategy	Use the h-last comments as context as-is.
Outcomes	+ LLMs are able to address and react to subtle textual cues such as phrasing, which are then passed down to the other users (App. B; Table 13). + Notable conversational events persist long after the context window elapses (App. B; Table 15). - Fine-grained information is lost if a user takes too long to respond (App. B; Table 15).

F Prompts Used

F.1 Personal user characteristics

We add random employment information to these personas, since we experimentally discovered that doing so improves discussion variety, and add a random subset of these personas to each discussion (§4.3). Each LLM participant is given a persona from the public version of the PERSONA

Table 19. Sociodemographic information provided to the LLM participants and annotators. We defer the reader to the supplementary material for the actual values.

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

dataset [16], since we were unable to establish contact with the authors. The parameters for each persona can be seen in Table 19. In order to promote greater variety in the LLM personas, we add a random trade whenever a persona is marked as being employed. The trades were generated using a GPT-5 model, and can be found in the Supplementary Material.

F.2 Participant Instructions

Below are the instructions for normal users and troll users respectively. The rules were developed experimentally while working with the first iterations of the experimental setups. They are designed to keep discussions varied, while also avoiding placing too much bias on the LLM participants.

F.2.1 Normal users.

Instructions.

- Write short, natural comments similar to those of typical forum users.
- Refer to others using @username.
- You may express strong opinions and frustration.
- You may react more strongly if you feel repeatedly provoked.
- You know that a moderator is present and can ban you.
- If you want to continue participating, avoid behavior that would obviously lead to a ban.

Output Format.

- Provide only your forum comment.
- If staying silent, output the empty string "".

F.2.2 Troll users.

Instructions.

- Write short, natural comments similar to those of typical forum users.
- Refer to others using @username.
- Post comments intended to annoy, frustrate, or provoke other users.
- Use sarcasm, exaggeration, or playful mockery.
- You may challenge other users' opinions and push their buttons.
- Maintain the illusion of being a normal user while trolling.

Output Format.

- Provide only your forum comment.
- If staying silent, output the empty string """".

F.2.3 No Instructions.

Output Format.

- Provide only your forum comment.
- If staying silent, output the empty string """".

F.3 Facilitation Strategies

F.3.1 Constructive communications.

General Rules.

- Respond only if absolutely necessary. Otherwise, remain silent.
- You are a human moderator in an online forum discussion.
- If intervention is necessary: output your short moderator message.
- If not: output ''.
- When responding, write concise, natural-sounding messages. Refer to users as @username.
- You may issue warnings. You may take disciplinary action (e.g., banning a user)

The Role of the Moderator.

- **Environment and Care:** Moderators create a warm environment, treating participants as human beings first.
- **Process Management:** Duties include timekeeping, monitoring energy levels, and managing physical needs like bio breaks or stretching.
- **Participation Balancing:** Mods must encourage shared 'air time,', normalize silence for thinking, and ensure participants talk to each other rather than through the moderator.
- **Conflict and Behavior:** The role involves highlighting points of disagreement to reach understanding and interrupting problematic behaviors like microaggressions or personal slights.

Limitations and Boundaries ('What NOT to do').

- **Non-Directive Approach:** Moderators are not decision-makers; they act as 'support staff' or 'tour managers' rather than conductors.
- **Deferring to participants:** Legitimacy lies with the participants, who set their own agenda and write/edit all content.
- **Content Neutrality:** Mods must not use their own knowledge of sustainability or fact-check participants, maintaining a strict separation between process and content.
- **Minimal Intervention:** Intervention is only permitted for reasons of respect, safety, or process clarification.

Tools for Effective Facilitation.

- **Active Listening:** Techniques include nodding, summarizing, and focusing on experiences rather than opinions.
- **Non-Content Prompts:** Helpful phrases include 'Tell me more,' 'What is the story behind that?', and 'Who hasn't had a chance to weigh in?'
- **Bias Awareness:** Moderators are encouraged to identify subconscious biases toward activists or administrators and to intentionally engage with ideas they may instinctively dislike.

F.3.2 Constructive communications.

General Rules.

- Respond only if absolutely necessary. Otherwise, remain silent.
- You are a human moderator in an online forum discussion.
- If intervention is necessary: output your short moderator message.
- If not: output "".
- When responding, write concise, natural-sounding messages. Refer to users as @username.
- You may issue warnings. You may take disciplinary action (e.g., banning a user)

Foundational Principles.

- Maintain Neutrality: Remain neutral at all times. Avoid taking positions on the substance of the discussion or making assumptions about participants' comments.
- Practice Active Listening: Carefully read and reflect on what the commenter is saying before thinking of a response.
- Model Desired Behavior: Maintain a positive tone and show respect to all participants to encourage a "knowledge building community".
- Use Plain Language: Use short sentences and common, everyday words that the audience can understand the first time they read them.

Engagement & Tone.

- Be Welcoming and Unique: Make participants feel appreciated and part of the community. Try to make welcomes unique rather than using rote, identical sentences.
- Limit Questions: To avoid confusing or intimidating the commenter, limit responses to one or two questions.
- Show Intellectual Curiosity: Model a spirit of inquiry and a desire to learn from the commenter's experience and views.

Facilitating Effective Comments.

- Mentor Effective Commenting: Move participants past "voting and venting" behaviors toward contributing substantive information they possess.
- Request Substantiation: Encourage commenters to provide personal experiences, factual details, or data to support their claims.
- Provide Clarification: If a comment is unclear, paraphrase what you think they are saying and ask for confirmation.
- Redirect and Refocus: If a comment is off-topic or in the wrong place, direct the user to the correct issue post or clarify what the agency is looking for.

Community Supervision & Technical Support.

- Police Civility: Intervene when interactions become heated, offensive, or involve personalized attacks.

F.3.3 No Instructions.

General Rules.

- Respond only if absolutely necessary. Otherwise, remain silent.
- You are a human moderator in an online forum discussion.
- If intervention is necessary: output your short moderator message.
- If not: output "".
- When responding, write concise, natural-sounding messages. Refer to users as @username.

- You may issue warnings. You may take disciplinary action (e.g., banning a user)

References

[1] Sahar Abdelnabi, Amr Gomaa, Sarath Sivaprasad, Lea Schönherr, and Mario Fritz. 2024. Cooperation, Competition, and Maliciousness: LLM-Stakeholders Interactive Negotiation. In *Advances in Neural Information Processing Systems*, A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (Eds.), Vol. 37. Curran Associates, Inc., 83548–83599. doi:10.52202/079017-2658

[2] Carlo Adornetto, Adrian Mora, Kai Hu, Leticia Izquierdo Garcia, Parfait Atchade-Adelomou, Gianluigi Greco, Luis Alberto Alonso Pastor, and Kent Larson. 2025. Generative Agents in Agent-Based Modeling: Overview, Validation, and Emerging Challenges. *IEEE Transactions on Artificial Intelligence* 6, 12 (2025), 3165–3183. doi:10.1109/TAI.2025.3566362

[3] T. Amaury and C. Stefano. 2022. Make Reddit Great Again: Assessing Community Effects of Moderation Interventions on r/The_Donald. *Proceedings of the ACM on Human-Computer Interaction* 6 (2022), 1 – 28. https://api.semanticscholar.org/CorpusID:246016404

[4] Lucas Anastasiou, Aldo De Moor, Barbara Brayshay, and Anna De Liddo. 2023. A tale of struggles: an evaluation framework for transitioning from individually usable to community-useful online deliberation tools. In *Proceedings of the 11th International Conference on Communities and Technologies* (Lahti, Finland) *(C&T '23)*. Association for Computing Machinery, New York, NY, USA, 144–155. doi:10.1145/3593743.3593771

[5] Jacy Reese Anthis, Ryan Liu, Sean M. Richardson, Austin C. Kozlowski, Bernard Koch, James Evans, Erik Brynjolfsson, and Michael Bernstein. 2025. LLM Social Simulations Are a Promising Research Method. arXiv:2504.02234 [cs.HC] https://arxiv.org/abs/2504.02234

[6] Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. 2024. Refusal in Language Models Is Mediated by a Single Direction. In *Advances in Neural Information Processing Systems*, A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (Eds.), Vol. 37. Curran Associates, Inc., 136037–136083. https://proceedings.neurips.cc/paper_files/paper/2024/file/f545448535dfde4f9786555403ab7c49-Paper-Conference.pdf

[7] 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 (2023), 1–8.

[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. arXiv:2408.04667 [cs.CL] https://arxiv.org/abs/2408.04667

[9] 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 (2024), 582 – 589. https://api.semanticscholar.org/CorpusID:268550181

[10] 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. arXiv:2408.08379 [cs.CL] https://arxiv.org/abs/2408.08379

[11] Rajiv D. Banker and Srikant M. Datar. 1989. Software Complexity and Maintainability. In *Proceedings of the 10th International Conference on Information Systems (ICIS)*. Association for Information Systems, AIS Electronic Library (AISeL), Boston, MA.

[12] Kent Beck. 2000. *Extreme programming explained: embrace change.* addison-wesley professional.

[13] James Bisbee, Joshua D. Clinton, Cassy Dorff, Brenton Kenkel, and Jennifer M. Larson. 2024. Synthetic Replacements for Human Survey Data? The Perils of Large Language Models. *Political Analysis* 32, 4 (2024), 401–416. doi:10.1017/pan.2024.5

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

[15] Kaitlin Emily Cannava. 2016. *The Repertoire of Understanding: The Linguistic Patterning of Repetition and Alignment within Supportive Conversations.* Louisiana State University and Agricultural & Mechanical College.

[16] Louis Castricato, Nathan Lile, Rafael Rafailov, Jan-Philipp Fränken, and Chelsea Finn. 2025. PERSONA: A Reproducible Testbed for Pluralistic Alignment. In *Proceedings of the 31st International Conference on Computational Linguistics*, Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert (Eds.). Association for Computational Linguistics, Abu Dhabi, UAE, 11348–11368. https://aclanthology.org/2025.coling-main.752/

[17] Jonathan P. Chang and Cristian Danescu. 2019. Trouble on the Horizon: Forecasting the Derailment of Online Conversations as they Develop. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Kentaro Inui,

Jing Jiang, Vincent Ng, and Xiaojun Wan (Eds.). Association for Computational Linguistics, Hong Kong, China, 4743–4754. doi:10.18653/v1/D19-1481

[18] Ming-Bin Chen, Lea Frermann, and Jey Han Lau. 2025. WHoW: A Cross-domain Approach for Analysing Conversation Moderation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, Luis Chiruzzo, Alan Ritter, and Lu Wang (Eds.). Association for Computational Linguistics, Albuquerque, New Mexico, 2091–2126. doi:10.18653/v1/2025.naacl-long.105

[19] Tuochao Chen, Nicholas Scott Batchelder, Alisa Liu, Noah A. Smith, and Shyamnath Gollakota. 2025. LlamaPIE: Proactive In-Ear Conversation Assistants. In *Findings of the Association for Computational Linguistics: ACL 2025*, Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (Eds.). Association for Computational Linguistics, Vienna, Austria, 13801–13824. doi:10.18653/v1/2025.findings-acl.710

[20] Yanran Chen, Jonas Belouadi, and Steffen Eger. 2022. Reproducibility Issues for BERT-based Evaluation Metrics. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (Eds.). Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 2965–2989. doi:10.18653/v1/2022.emnlp-main.192

[21] Hyundong Cho, Shuai Liu, Taiwei Shi, Darpan Jain, Basem Rizk, Yuyang Huang, Zixun Lu, Nuan Wen, Jonathan Gratch, Emilio Ferrara, and Jonathan 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)*, Kevin Duh, Helena Gomez, and Steven Bethard (Eds.). Association for Computational Linguistics, Mexico City, Mexico, 7478–7496. doi:10.18653/v1/2024.naacl-long.415

[22] Yun-Shiuan Chuang, Agam Goyal, Nikunj Harlalka, Siddharth Suresh, Robert Hawkins, Sijia Yang, Dhavan Shah, Junjie Hu, and Timothy Rogers. 2024. Simulating Opinion Dynamics with Networks of LLM-based Agents. In *Findings of the Association for Computational Linguistics: NAACL 2024*, Kevin Duh, Helena Gomez, and Steven Bethard (Eds.). Association for Computational Linguistics, Mexico City, Mexico, 3326–3346. doi:10.18653/v1/2024.findings-naacl.211

[23] Erik K. Clemons, Matthias Schreieck, and Rishi V. Waran. 2025. Managing disinformation on social media platforms. *Electronic Markets* 35, 52 (2025). doi:10.1007/s12525-025-00796-6

[24] Bryn Alexander Coles and Melanie West. 2016. Trolling the trolls: Online forum users constructions of the nature and properties of trolling. *Computers in Human Behavior* 60 (2016), 233–244.

[25] Andrew Collins, Matthew Koehler, and Christopher Lynch. 2024. Methods That Support the Validation of Agent-Based Models: An Overview and Discussion. *Journal of Artificial Societies and Social Simulation* 27, 1 (2024), 11. doi:10.18564/jasss.5258

[26] 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* (Barcelona, Spain) *(HT '22)*. Association for Computing Machinery, New York, NY, USA, 248–251. doi:10.1145/3511095.3536369

[27] E. Denniss and R. Lindberg. 2025. Social media and the spread of misinformation: infectious and a threat to public health. *Health Promotion International* 40 (2025). doi:10.1093/heapro/daaf023

[28] Joshua M. Epstein. 1999. Agent-based computational models and generative social science. *Complex.* 4, 5 (May 1999), 41–60.

[29] Cornell eRulemaking Initiative. 2017. CeRI (Cornell e-Rulemaking) Moderator Protocol. https://scholarship.law.cornell.edu/ceri/21 Cornell e-Rulemaking Initiative Publications, 21.

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

[31] Davide Falessi, Lionel C. Briand, Giovanni Cantone, Rafael Capilla, and Philippe Kruchten. 2013. The value of design rationale information. *ACM Trans. Softw. Eng. Methodol.* 22, 3, Article 21 (July 2013), 32 pages. doi:10.1145/2491509.2491515

[32] 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*, Khalid Al-Khatib, Yufang Hou, and Manfred Stede (Eds.). Association for Computational Linguistics, Punta Cana, Dominican Republic, 133–141. doi:10.18653/v1/2021.argmining-1.13

[33] Neele Falk, Eva Vecchi, Iman Jundi, and Gabriella Lapesa. 2024. Moderation in the Wild: Investigating User-Driven Moderation in Online Discussions. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, Yvette Graham and Matthew Purver (Eds.). Association for Computational Linguistics, St. Julian's, Malta, 992–1013. https://aclanthology.org/2024.eacl-long.60/

[34] Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed. 2024. Bias and Fairness in Large Language Models: A Survey. *Computational Linguistics* 50, 3 (Sept. 2024), 1097–1179. doi:10.1162/coli_a_00524

[35] Chen Gao, Xiaochong Lan, Zhihong Lu, Jinzhu Mao, Jinghua Piao, Huandong Wang, Depeng Jin, and Yong Li. 2025. S^3: Social-network Simulation System with Large Language Model-Empowered Agents. arXiv:2307.14984 [cs.SI] https://arxiv.org/abs/2307.14984

[36] Rena Gao, Ming-Bin Chen, Lea Frermann, and Jey Han Lau. 2025. Moderation Matters: Measuring Conversational Moderation Impact in English as a Second Language Group Discussion. In *Findings of the Association for Computational Linguistics: ACL 2025*, Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (Eds.). Association for Computational Linguistics, Vienna, Austria, 2070–2095. doi:10.18653/v1/2025.findings-acl.106

[37] Kristina Gligori'c, Tijana Zrnic, Cinoo Lee, Emmanuel J. Candes, and Dan Jurafsky. 2024. Can Unconfident LLM Annotations Be Used for Confident Conclusions? *ArXiv* abs/2408.15204 (2024). https://api.semanticscholar.org/CorpusID:271962879

[38] Nitesh Goyal, Ian D. Kivlichan, Rachel Rosen, and Lucy Vasserman. 2022. Is Your Toxicity My Toxicity? Exploring the Impact of Rater Identity on Toxicity Annotation. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW2, Article 363 (Nov. 2022), 28 pages. doi:10.1145/3555088

[39] Igor Grossmann, Matthew Feinberg, Dawn Parker, Nicholas Christakis, Philip Tetlock, and William Cunningham. 2023. AI and the transformation of social science research. *Science (New York, N.Y.)* 380 (06 2023), 1108–1109. doi:10.1126/science.adi1778

[40] Mahyar Habibi, Dirk Hovy, and Carlo Schwarz. 2025. The Content Moderator's Dilemma: Removal of Toxic Content and Distortions to Online Discourse. arXiv:2412.16114 [cs.SI] https://arxiv.org/abs/2412.16114

[41] Michael Hanna and Ondřej Bojar. 2021. A Fine-Grained Analysis of BERTScore. In *Proceedings of the Sixth Conference on Machine Translation*, Loic Barrault, Ondrej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussa, Christian Federmann, Mark Fishel, Alexander Fraser, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Tom Kocmi, Andre Martins, Makoto Morishita, and Christof Monz (Eds.). Association for Computational Linguistics, Online, 507–517. https://aclanthology.org/2021.wmt-1.59/

[42] Luke Hewitt, Ashwini Ashokkumar, Isaias Ghezae, and Robb Willer. 2024. Predicting Results of Social Science Experiments Using Large Language Models. (08 2024). Equal contribution, order randomized.

[43] Christopher Hopkinson. 2013. Trolling in online discussions: From provocation to community-building. *Brno studies in English* 39, 1 (2013), 5–25.

[44] 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* (Austin, TX, USA) *(WWW '23)*. Association for Computing Machinery, New York, NY, USA, 2666–2676. doi:10.1145/3543507.3583275

[45] Wei Huang, Xingyu Zheng, Xudong Ma, Haotong Qin, Chengtao Lv, Hong Chen, Jie Luo, Xiaojuan Qi, Xianglong Liu, and Michele Magno. 2024. An empirical study of LLaMA3 quantization: from LLMs to MLLMs. *Visual Intelligence* 2, 1 (30 Dec 2024), 36. doi:10.1007/s44267-024-00070-x

[46] Guido Ivetta, Laura Moradbakhti, and Rafael A Calvo. 2025. Synthetic social data: trials and tribulations. *arXiv preprint arXiv:2510.19952* (2025).

[47] Andrew Jakubowicz, Kevin Dunn, Gail Mason, Yin Paradies, Ana-Maria Bliuc, Nasya Bahfen, Andre Oboler, Rosalie Atie, and Karen Connelly. 2017. *Cyber racism and community resilience.* Springer.

[48] Bernard J. Jansen, Soon gyo Jung, and Joni Salminen. 2023. Employing large language models in survey research. *Natural Language Processing Journal* 4 (2023), 100020. doi:10.1016/j.nlp.2023.100020

[49] Carolin Kaiser, Jakob Kaiser, Vladimir Manewitsch, Lea Rau, and Rene Schallner. 2025. Simulating human opinions with large language models: Opportunities and challenges for personalized survey data modeling. In *Adjunct Proceedings of the 33rd ACM Conference on User Modeling, Adaptation and Personalization.* 82–86.

[50] 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, Article 87 (April 2021), 26 pages. doi:10.1145/3449161

[51] 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 (June 2025). doi:10.1177/20539517251321752

[52] Katerina Korre, Dimitris Tsirmpas, Nikos Gkoumas, Emma Cabalé, Danai Myrtzani, Theodoros Evgeniou, Ion Androutsopoulos, and John Pavlopoulos. 2025. Evaluation and Facilitation of Online Discussions in the LLM Era: A Survey. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (Eds.). Association for Computational Linguistics, Suzhou, China, 24443–24462. doi:10.18653/v1/2025.emnlp-main.1243

[53] Sarah Kreps and Doug Kriner. 2023. How AI threatens democracy. *Journal of Democracy* 34, 4 (2023), 122–131.

[54] Michael Krumdick, Charles Lovering, Varshini Reddy, Seth Ebner, and Chris Tanner. 2025. No free labels: Limitations of llm-as-a-judge without human grounding. *arXiv preprint arXiv:2503.05061* (2025).

[55] 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 (2024), 865–878.

[56] Eldar Kurtic, Alexandre Noll Marques, Shubhra Pandit, Mark Kurtz, and Dan Alistarh. 2025. "Give Me BF16 or Give Me Death"? Accuracy-Performance Trade-Offs in LLM Quantization. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (Eds.). Association for Computational Linguistics, Vienna, Austria, 26872–26886. doi:10.18653/v1/2025.acl-long.1304

[57] Ang Li, Yin Zhou, Vethavikashini Chithrra Raghuram, Tom Goldstein, and Micah Goldblum. 2025. Commercial LLM Agents Are Already Vulnerable to Simple Yet Dangerous Attacks. arXiv:2502.08586 [cs.LG] https://arxiv.org/abs/2502.08586

[58] Nian Li, Chen Gao, Mingyu Li, Yong Li, and Qingmin Liao. 2024. EconAgent: Large Language Model-Empowered Agents for Simulating Macroeconomic Activities. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 15523–15536. doi:10.18653/v1/2024.acl-long.829

[59] Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*. Association for Computational Linguistics, Barcelona, Spain, 74–81. https://aclanthology.org/W04-1013

[60] Yi Liu, Gelei Deng, Zhengzi Xu, Yuekang Li, Yaowen Zheng, Ying Zhang, Lida Zhao, Tianwei Zhang, Kailong Wang, and Yang Liu. 2023. Jailbreaking chatgpt via prompt engineering: An empirical study. *arXiv preprint arXiv:2305.13860* (2023).

[61] Yiran Liu, Ke Yang, Zehan Qi, Xiao Liu, Yang Yu, and ChengXiang Zhai. 2024. Bias and Volatility: A Statistical Framework for Evaluating Large Language Model's Stereotypes and the Associated Generation Inconsistency. In *Advances in Neural Information Processing Systems*, A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (Eds.), Vol. 37. Curran Associates, Inc., 110131–110155. doi:10.52202/079017-3495

[62] Emma Rose Madden. 2025. Evaluating the Use of Large Language Models as Synthetic Social Agents in Social Science Research. *Journal of Social Computing* 6, 4 (2025), 334–341.

[63] Durjoy Majumdar, Arjun S, Pranavi Boyina, Sri Sai Priya Rayidi, Yerra Rahul Sai, and Suryakanth V Gangashetty. 2024. Beyond Text: Nefarious Actors Harnessing LLMs for Strategic Advantage. In *2024 International Conference on Intelligent Systems for Cybersecurity (ISCS)*. 1–7. doi:10.1109/ISCS61804.2024.10581181

[64] Evita March and Jessica Marrington. 2019. A Qualitative Analysis of Internet Trolling. *Cyberpsychology, Behavior, and Social Networking* 22, 3 (2019), 192–197. arXiv:https://doi.org/10.1089/cyber.2018.0210 doi:10.1089/cyber.2018.0210 PMID: 30720370.

[65] Fiammetta Marulli, Pierluigi Paganini, and Fabio Lancellotti. 2024. The Three Sides of the Moon LLMs in Cybersecurity: Guardians, Enablers and Targets. *Procedia Computer Science* 246 (2024), 5340–5348. doi:10.1016/j.procs.2024.09.653 28th International Conference on Knowledge Based and Intelligent information and Engineering Systems (KES 2024).

[66] MIT Center for Constructive Communication. 2024. Unpublished training materials developed by the MIT Center for Constructive Communication. (2024). Guide given to human facilitators.

[67] Xinyi Mou, Zhongyu Wei, and Xuanjing Huang. 2024. Unveiling the Truth and Facilitating Change: Towards Agent-based Large-scale Social Movement Simulation. In *Findings of the Association for Computational Linguistics: ACL 2024*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 4789–4809. doi:10.18653/v1/2024.findings-acl.285

[68] Alejandro Leonardo García Navarro, Nataliia Koneva, Alfonso Sánchez-Macián, José Alberto Hernández, and Manuel Goyanes. 2024. Designing Reliable Experiments with Generative Agent-Based Modeling: A Comprehensive Guide Using Concordia by Google DeepMind. arXiv:2411.07038 [cs.AI] https://arxiv.org/abs/2411.07038

[69] Terrence Neumann, Maria De-Arteaga, and Sina Fazelpour. 2025. Should you use LLMs to simulate opinions? Quality checks for early-stage deliberation. arXiv:2504.08954 [cs.CY] https://arxiv.org/abs/2504.08954

[70] Ryota Nonomura and Hiroki Mori. 2025. Who speaks next? Multi-party AI discussion leveraging the systematics of turn-taking in Murder Mystery games. *Frontiers in Artificial Intelligence* Volume 8 - 2025 (2025). doi:10.3389/frai.2025.1582287

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

[72] OpenAI. 2024. GPT-4 Technical Report. arXiv:2303.08774 [cs.CL] https://arxiv.org/abs/2303.08774

[73] OpenAI Pricing 2026. *Open AI pricing*. Retrieved 2026-03-05 from https://openai.com/api/pricing/

[74] Guanzhong Pan, Vishal Chodnekar, Abinas Roy, and Haibo Wang. 2025. A Cost-Benefit Analysis of On-Premise Large Language Model Deployment: Breaking Even with Commercial LLM Services. arXiv:2509.18101 [cs.AI] https://arxiv.org/abs/2509.18101

[75] J. Park, S. Klingel, C. Cardie, M. Newhart, C. Farina, and J.J. Vallbé. 2012. Facilitative moderation for online participation in eRulemaking. In *Proceedings of the 13th Annual International Conference on Digital Government Research* (College Park, Maryland, USA). New York, NY, USA, 173–182. doi:10.1145/2307729.2307757

[76] 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. arXiv:2304.03442 [cs.HC] https://arxiv.org/abs/2304.03442

[77] 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* (Bend, OR, USA) *(UIST '22)*. Association for Computing Machinery, New York, NY, USA, Article 74, 18 pages. doi:10.1145/3526113.3545616

[78] Joon Sung Park, Carolyn Q. Zou, Aaron Shaw, Benjamin Mako Hill, Carrie Cai, Meredith Ringel Morris, Robb Willer, Percy Liang, and Michael S. Bernstein. 2024. Generative Agent Simulations of 1,000 People. arXiv:2411.10109 [cs.AI] https://arxiv.org/abs/2411.10109

[79] 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*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (Eds.). Association for Computational Linguistics, Online, 4296–4305. doi:10.18653/v1/2020.acl-main.396

[80] Perspective API 2026. *Perspective API*. https://www.perspectiveapi.com/ Accessed: 2026-02-24.

[81] Matt Post. 2018. A Call for Clarity in Reporting BLEU Scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana Neves, Matt Post, Lucia Specia, Marco Turchi, and Karin Verspoor (Eds.). Association for Computational Linguistics, Brussels, Belgium, 186–191. doi:10.18653/v1/W18-6319

[82] Luiza Pozzobon, Beyza Ermis, Patrick Lewis, and Sara Hooker. 2023. On the Challenges of Using Black-Box APIs for Toxicity Evaluation in Research. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 7595–7609. doi:10.18653/v1/2023.emnlp-main.472

[83] FRANZISKA PRADEL, JAN ZILINSKY, SPYROS KOSMIDIS, and YANNIS THEOCHARIS. 2024. Toxic Speech and Limited Demand for Content Moderation on Social Media. *American Political Science Review* 118, 4 (2024), 1895–1912. doi:10.1017/S000305542300134X

[84] Samana Pranesh and Sayan Gupta. 2024. The impact of social media on polarization in the society. arXiv:2408.12877 [physics.soc-ph] https://arxiv.org/abs/2408.12877

[85] William Rand and Roland T. Rust. 2011. Agent-based modeling in marketing: Guidelines for rigor. *International Journal of Research in Marketing* 28, 3 (2011), 181–193. doi:10.1016/j.ijresmar.2011.04.002

[86] Craig W. Reynolds. 1998. *Flocks, herds, and schools: a distributed behavioral model*. Association for Computing Machinery, New York, NY, USA, 273–282. https://doi.org/10.1145/280811.281008

[87] 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. arXiv:2408.00818 [cs.AI] https://arxiv.org/abs/2408.00818

[88] Giovanni Rossi. 2020. Other-repetition in conversation across languages: Bringing prosody into pragmatic typology. *Language in Society* 49, 4 (2020), 495–520.

[89] Luca Rossi, Katherine Harrison, and Irina Shklovski. 2024. The Problems of LLM-generated Data in Social Science Research. *Sociologica* 18, 2 (Jan. 2024), 145–168. doi:10.6092/issn.1971-8853/19576

[90] Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose opinions do language models reflect?. In *Proceedings of the 40th International Conference on Machine Learning* (Honolulu, Hawaii, USA) *(ICML'23)*. JMLR.org, Article 1244, 34 pages.

[91] Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A. Smith. 2022. Annotators with Attitudes: How Annotator Beliefs And Identities Bias Toxic Language Detection. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz (Eds.). Association for Computational Linguistics, Seattle, United States, 5884–5906. doi:10.18653/v1/2022.naacl-main.431

[92] Carlos Saura García. 2025. Synthetification of public opinion: impacts on deliberative democracies. *Ethics and Information Technology* 27, 4 (2025), 54.

[93] 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* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing

Machinery, New York, NY, USA, Article 486, 16 pages. doi:10.1145/3613904.3642333

[94] Thomas C. Schelling. 1971. Dynamic models of segregation†. *The Journal of Mathematical Sociology* 1, 2 (1971), 143–186. arXiv:https://doi.org/10.1080/0022250X.1971.9989794 doi:10.1080/0022250X.1971.9989794

[95] 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, Article 370 (Nov. 2022), 27 pages. doi:10.1145/3555095

[96] Angela M Schöpke-Gonzalez, Shubham Atreja, Han Na Shin, Najmin Ahmed, and Libby Hemphill. 2024. Why do volunteer content moderators quit? Burnout, conflict, and harmful behaviors. *New Media & Society* 26, 10 (2024), 5677–5701.

[97] 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*. Bangkok, Thailand, 13985–14001.

[98] 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, Article 107 (Oct. 2020), 28 pages. doi:10.1145/3415178

[99] 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 (2023). https://api.semanticscholar.org/CorpusID:259211996

[100] Chandler Smith, Rakshit Trivedi, Jesse Clifton, Lewis Hammond, Akbir Khan, Sasha Vezhnevets, John P Agapiou, Edgar A. Duéñez-Guzmán, Jayd Matyas, Danny Karmon, Marwa Abdulhai, Dylan Hadfield-Menell, Natasha Jaques, Joel Z Leibo, Oliver Slumbers, Tim Baarslag, and Minsuk Chang. 2024. The Concordia Contest: Advancing the Cooperative Intelligence of Language Agents. In *NeurIPS 2024 Competition Track*. https://openreview.net/forum?id=dfeFy1PSSw

[101] Pia Sommerauer, Giulia Rambelli, and Tommaso Caselli. 2025. Simulating Identity, Propagating Bias: Abstraction and Stereotypes in LLM-Generated Text. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (Eds.). Association for Computational Linguistics, Suzhou, China, 19812–19831. doi:10.18653/v1/2025.findings-emnlp.1080

[102] Ruth Spence, Antonia Bifulco, Paula Bradbury, Elena Martellozzo, and Jeffrey DeMarco. 2023. The psychological impacts of content moderation on content moderators: A qualitative study. *Cyberpsychology: Journal of Psychosocial Research on Cyberspace* 17, 4 (2023).

[103] Ruth Spence, Antonia Bifulco, Paula Bradbury, Elena Martellozzo, and Jeffrey DeMarco. 2024. Content moderator mental health, secondary trauma, and well-being: A cross-sectional study. *Cyberpsychology, Behavior, and Social Networking* 27, 2 (2024), 149–155.

[104] Miriah Steiger, Timir J Bharucha, Sukrit Venkatagiri, Martin J Riedl, and Matthew Lease. 2021. The psychological well-being of content moderators: the emotional labor of commercial moderation and avenues for improving support. In *Proceedings of the 2021 CHI conference on human factors in computing systems*. 1–14.

[105] Christopher Summerfield, Lisa Argyle, Michiel Bakker, Teddy Collins, Esin Durmus, Tyna Eloundou, Iason Gabriel, Deep Ganguli, Kobi Hackenburg, Gillian Hadfield, et al. 2024. How will advanced AI systems impact democracy? *arXiv preprint arXiv:2409.06729* (2024).

[106] Qiusi Sun and Cuihua Shen. 2021. Who would respond to A troll? A social network analysis of reactions to trolls in online communities. *Computers in Human Behavior* 121 (2021), 106786.

[107] Jan Svennevig. 2004. Other-repetition as display of hearing, understanding and emotional stance. *Discourse studies* 6, 4 (2004), 489–516.

[108] Amir Taubenfeld, Yaniv Dover, Roi Reichart, and Ariel Goldstein. 2024. Systematic Biases in LLM Simulations of Debates. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 251–267. doi:10.18653/v1/2024.emnlp-main.16

[109] Gemini Team. 2025. Gemini: A Family of Highly Capable Multimodal Models. arXiv:2312.11805 [cs.CL] https://arxiv.org/abs/2312.11805

[110] Dave Thomas. 2025. *simplicity: sustainable, humane, and effective software development.* The Pragmatic Programmers LLC.

[111] Lily L. Tsai, Alex Pentland, Alia Braley, Nuole Chen, José Ramón Enríquez, and Anka Reuel. 2024. *Generative AI for Pro-Democracy Platforms.* https://mit-genai.pubpub.org/pub/mn45hexw.

[112] Vivian Tsai, Crystal Qian, and Deliberate Lab comm. contributors. 2025. Deliberate Lab: Open-Source Platform for LLM-Powered Social Science. https://github.com/PAIR-code/deliberate-lab

[113] Dimitris Tsirmpas. 2025. *PEFK: Patterns of Engagement in Facilitated Konversations.* https://github.com/dimits-ts/facilitation-dataset Dataset. Released 2025-06-19.

[114] Dimitrios Tsirmpas, Ioannis Gkionis, Georgios Th. Papadopoulos, and Ioannis Mademlis. 2024. Neural natural language processing for long texts: A survey on classification and summarization. *Engineering Applications of Artificial Intelligence* 133 (2024), 108231. doi:10.1016/j.engappai.2024.108231

[115] Petter Törnberg, Diliara Valeeva, Justus Uitermark, and Christopher Bail. 2023. Simulating Social Media Using Large Language Models to Evaluate Alternative News Feed Algorithms. arXiv:2310.05984 [cs.SI] https://arxiv.org/abs/2310.05984

[116] 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 (2024). https://api.semanticscholar.org/CorpusID:266902624

[117] 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 (2023). https://api.semanticscholar.org/CorpusID:265692435

[118] Alexander Sasha Vezhnevets, John P. Agapiou, Avia Aharon, Ron Ziv, Jayd Matyas, Edgar A. Duéñez-Guzmán, William A. Cunningham, Simon Osindero, Danny Karmon, and Joel Z. Leibo. 2023. Generative agent-based modeling with actions grounded in physical, social, or digital space using Concordia. arXiv:2312.03664 [cs.AI] https://arxiv.org/abs/2312.03664

[119] Huandong Wang, Wenjie Fu, Yingzhou Tang, Zhilong Chen, Yuxi Huang, Jinghua Piao, Chen Gao, Fengli Xu, Tao Jiang, and Yong Li. 2025. A Survey on Responsible LLMs: Inherent Risk, Malicious Use, and Mitigation Strategy. arXiv:2501.09431 [cs.AI] https://arxiv.org/abs/2501.09431

[120] Mark Warner, Angelika Strohmayer, Matthew Higgs, and Lynne Coventry. 2025. A critical reflection on the use of toxicity detection algorithms in proactive content moderation systems. *International Journal of Human-Computer Studies* 198 (2025), 103468. doi:10.1016/j.ijhcs.2025.103468

[121] Zengqing Wu, Run Peng, Takayuki Ito, Makoto Onizuka, and Chuan Xiao. 2025. Llm-based social simulations require a boundary. *arXiv preprint arXiv:2506.19806* (2025).

[122] Yukun Yang. 2019. When power goes wild online: How did a voluntary moderator's abuse of power affect an online community? *Proceedings of the Association for Information Science and Technology* 56, 1 (2019), 504–508.

[123] Sibo Yi, Yule Liu, Zhen Sun, Tianshuo Cong, Xinlei He, Jiaxing Song, Ke Xu, and Qi Li. 2024. Jailbreak Attacks and Defenses Against Large Language Models: A Survey. arXiv:2407.04295 [cs.CR] https://arxiv.org/abs/2407.04295

[124] Richard J. Young. 2026. Comparative Analysis of LLM Abliteration Methods: A Cross-Architecture Evaluation. arXiv:2512.13655 [cs.CL] https://arxiv.org/abs/2512.13655

[125] 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. arXiv:2407.06567 [cs.CL]

[126] Gaofeng Yue, Jihong Liu, Qiang Zhang, and Yongzhu Hou. 2023. Building a Design-Rationale-Centric Knowledge Network to Realize the Internalization of Explicit Knowledge. *Applied Sciences* 13, 3 (2023). doi:10.3390/app13031539

[127] Ruoyu Zhang, Yanzeng Li, Yongliang Ma, Ming Zhou, and Lei Zou. 2023. Llmaaa: Making large language models as active annotators. In *Findings of the Association for Computational Linguistics: EMNLP 2023*. 13088–13103.

[128] Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, and Maarten Sap. 2024. SOTOPIA: Interactive Evaluation for Social Intelligence in Language Agents. In *The Twelfth International Conference on Learning Representations*. https://openreview.net/forum?id=mM7VurbA4r