Athens University of Economics and Business

Information Processing Laboratory
Natural Language Processing Group
Athens, Greece

Master Thesis
in
Data Science

Mitigating Polarisation in Online Discussions Through Adaptive Moderation Techniques

Dimitris Tsirmpas

Supervisor: Assistant Professor John Pavlopoulos
Department of Informatics
Athens University of Economics and Business

Committee: Director Vasilios Vassalos
Department of Informatics
Information Processing Laboratory
Athens University of Economics and Business

Associate Professor Themos Stafylakis
Department of Informatics
Athens University of Economics and Business

October 2024

Dimitris Tsirmpas

Mitigating Polarisation in Online Discussions Through Adaptive Moderation Techniques

October 2024

Supervisor: Assistant Professor John Pavlopoulos

Reviewers: Vasilios Vassalos, Themos Stafylakis

Athens University of Economics and Business

Information Processing Laboratory

Natural Language Processing Group

Department of Informatics

Athens, Greece Patision 76

10434 and Athens, Greece

Abstract

Online discussion moderation/facilitation is crucial for discussions to flourish and prevent polarization and toxicity, which nowadays seems omnipresent. However, because it is heavily based on humans, moderation/facilitation proves costly, time-consuming and non-scalable, which has led many to turn to LLMs for discourse facilitation. In this thesis, we explore the use of LLMs as pseudo-users in online discussions, as a cost-efficient, realistic and scalable way of substituting initial LLM facilitation experiments, which would ordinarily necessitate costly human involvement. Furthermore, we show that including socio-demographic backgrounds in our LLM users leads to more realistic discussions. We explore the use of LLM annotators to estimate discussion quality, using a new statistical test to gauge annotator polarization, and show that using socio-demographic backgrounds in LLM annotators may not meaningfully affect their judgments. Finally, we release a synthetic-discussion creation and annotation framework, the synthetic datasets resulting from our experiments, as well as subsequent analysis and findings from these datasets. Code, datasets and analysis can be found at `https://github.com/dimits-ts/llm_moderation_research`.

⚠Content Warning: This paper contains samples of harmful text, including violent, toxic, controversial, and potentially illegal statements.

Περίληψη

Ο συντονισμός/διαμεσολάβση (moderation/facilitation) των διαδικτυακών συζητήσεων είναι ζωτικής σημασίας για την άνθηση των συζητήσεων και την αποτροπή της πόλωσης και της τοξικότητας, που στις μέρες μας φαίνεται πανταχού παρούσες. Οι σύγχρονες τεχνικές συντονισμού/διαμεσολάβησης απαιτούν ανθρώπινη συμμετοχή και, ως εκ τούτου, είναι δαπανηρές και μη επεκτάσιμες, οδηγώντας πολλούς να στραφούν στη χρήση Μεγάλων Γλωσσικών Μοντέλων (ΜΓΜ, ή LLMs στα Αγγλικά) για αυτές. Στα πλαίσια της διατριβής αυτής δημιουργούμε ένα νέο σύστημα το οποίο παράγει συνθετικές διαδικτυακές συζητήσεις, χρησιμοποιώντας ψευτο-χρήστες ΜΓΜ με κοινωνικο-δημογραφικά υπόβαθρα έτσι ώστε να καταστήσουμε τις συζητήσεις ρεαλιστικές. Επιπλέον, δείχνουμε ότι η χρήση κοινωνικο-δημογραφικών υποβάθρων οδηγεί σε πιο ρεαλιστικές συζητήσεις. Διερευνούμε τη χρήση των σχολιαστών LLM για την εκτίμηση της ποιότητας των συζητήσεων, χρησιμοποιώντας ένα νέο στατιστικό έλεγχο για τη μέτρηση της πόλωσης των σχολιαστών και δείχνουμε ότι η χρήση κοινωνικο-δημογραφικού υπόβαθρου στους σχολιαστές LLM μπορεί να μην επηρεάζει σημαντικά τις κρίσεις τους. Επεκτείνουμε το σύστημα μας με τη δυνατότητα υποστήριξης αυτόματων επισημειωτών (με χρήση ΜΓΜ), για την αντιμετώπιση του προβλήματος της αξιολόγησης διαλόγων. Οι ψευτο-επισημειωτές αυτοί έχουν προκαθορισμένα από εμάς κοινωνικο-δημογραφικά υπόβαθρα, έτσι ώστε να προσομοιώσουμε τη διαφωνία που πιθανώς να υπάρχει ανάμεσα σε ανθρώπους με αντίστοιχα υπόβαθρα. Τέλος, δίνουμε στη δημοσιότητα το δικό μας πρόγραμμα δημιουργίας και σχολιασμού συνθετικών συζητήσεων, τα συνθετικά σύνολα δεδομένων που προέκυψαν από τα πειράματά μας, καθώς και την επακόλουθη ανάλυση και τα συμπεράσματα από αυτά. Ο κώδικας, τα σύνολα δεδομένων και η ανάλυση βρίσκονται στο αποθετήριο κώδικα στη διεύθυνση `https://github.com/dimits-ts/llm_moderation_research`.

Acknowledgements

I would like to thank Assistant Professor John Pavlopoulos for personally supervising the thesis and encouraging me to bring it to fruition until the very end. I also deeply thank Professor Ion Androutsopoulos for providing an invaluable, in-depth review during the development of the thesis. Both professors played an instrumental role in the final shape of this project from the very beginning, and played a key role in coordinating it with the rest of the research team. I would also like to thank the colleagues and researchers at the Archimedes/Athena Research Center, whose input helped steer the project towards more productive directions. Lastly, I am grateful to my family, whose constant support has greatly aided me in all my endeavors.

Contents

1 Introduction

1.1 Motivation and Problem Statement

Large Language Models (LLMs) have revolutionized the field of Natural Language Processing (NLP), especially after the introduction of ChatGPT in 2022. Their ability to not only convincingly produce human-like text, but also to respond to user queries and execute tasks such as summarization, annotation and classification [Tsi+24; Tan+24], have led many established companies, startups and research groups around the globe to scramble and identify useful use-cases for this novel technology [Had+23; Zho+24; Hut+24].

One such identified case is their use in online discussions. The online environment is essential to healthy democratic discourse [WS07; JK05; Pap04] and deliberative discussions [Sma+21], whose goal is for citizens to share opinions in order to make informed decisions. However, because of the anonymity online discussions offer [Ava+24], they are often characterized by aggression and toxicity [Xia+20], which often leads to low-quality discourse [WS07] (although the latter position is contested [Pap04]). Thus, discussions are often overseen by *"discourse moderators"*, people whose responsibility is to uphold the rules of the discussion and discipline users. In more formal environments *"discourse facilitators"* may be present, ensuring equal participation and helping the participants coordinate with one another. Other equally essential parts of facilitation are promoting even participation, dynamically summarizing the discussion, encouraging the sharing of ideas and opinions, and keeping discussions on-point [Har24; Wan08].

Nevertheless, human facilitation is expensive, time-consuming and often relies on specialized staff [Sma+23]. LLMs are perfectly positioned to aid in facilitating discussions [Sma+23], since they are relatively inexpensive, can be scaled easily, and their summarization and text-generation abilities are ideal for the facilitation tasks we outlined above. However, finding the correct prompts and configurations (e.g. which model family, whether to use pretrained or finetuned models, ...) by use of robust experiments with human subjects can be very difficult, and similarly expensive on the researchers' side, because of the heavy use of human participants. This effort represents the wider research context within which this thesis exists, and is illustrated in Figure 1.1.

In this thesis, we aim to address this limitation by leveraging LLMs to generate synthetic online discussions at scale. We develop a framework that can automatically produce synthetic

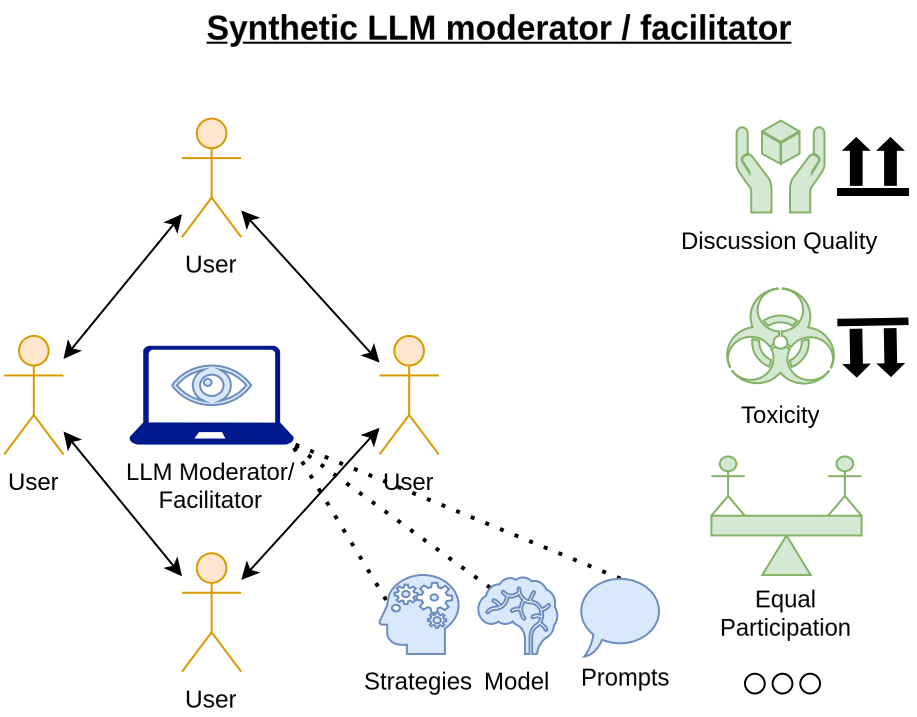

Fig. 1.1: The goal of the wider research context of this thesis: the selection of LLMs, moderation/facilitation strategies and the development of LLM prompts as to qualitatively improve online discussions.

discussions at scale, involving users with diverse Socio-Demographic Backgrounds (SDBs) at relatively low cost and within reasonable time constraints. The ability to generate these synthetic discussions easily, offers opportunities for low-cost experimentation, prototyping, and A/B testing. Additionally, the creation of a large synthetic dataset has potential applications for large-scale data analysis. In the context of prompt engineering, this effort can be seen as an adversarial procedure where some of the LLM user-agents try to derail the discussion, while the LLM moderator/facilitator attempts to keep it civil (Figure 1.2).

Our framework further incorporates automated LLM-based annotations of these synthetic discussions, allowing for an inexpensive comparison of the effects of various factors such as moderator strategy, moderator presence, and LLM user prompts. Ordinarily, using LLMs for annotation presents two distinct issues: the model's inherent biases and the question of how representative their annotations are in comparison with ones that would be made by humans. While the latter concern can only be conclusively addressed by a correlation study, we attempt to address the former by using annotators with different SDBs (Figure 1.3). This also allows us to assess whether and how different LLM personalities influence the annotation process.

Having set up our framework, we experiment with various prompt strategies and configurations to evaluate how they affect discussion quality, using toxicity as a proxy. Finally, we analyze the content of the discussions alongside the LLM annotations and generate

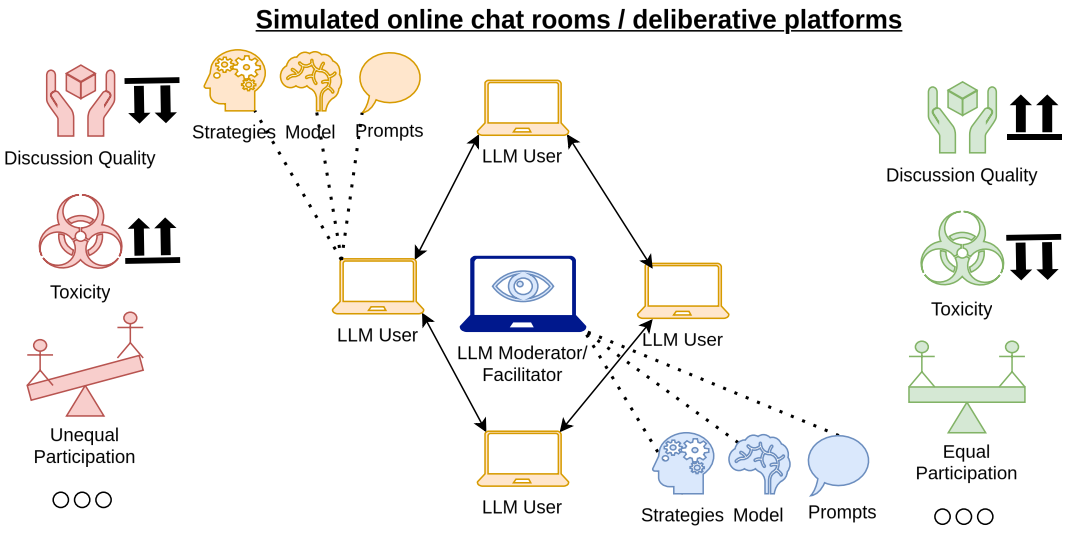

Fig. 1.2: The subject of this thesis; developing a framework where many LLM user-agents can simulate online discussions. We prime the LLM user-agents to uphold their personal opinions, even if doing so lowers the quality of the discussion. At the same time, we instruct the LLM-moderator/facilitator to keep the discussion quality as high as possible.

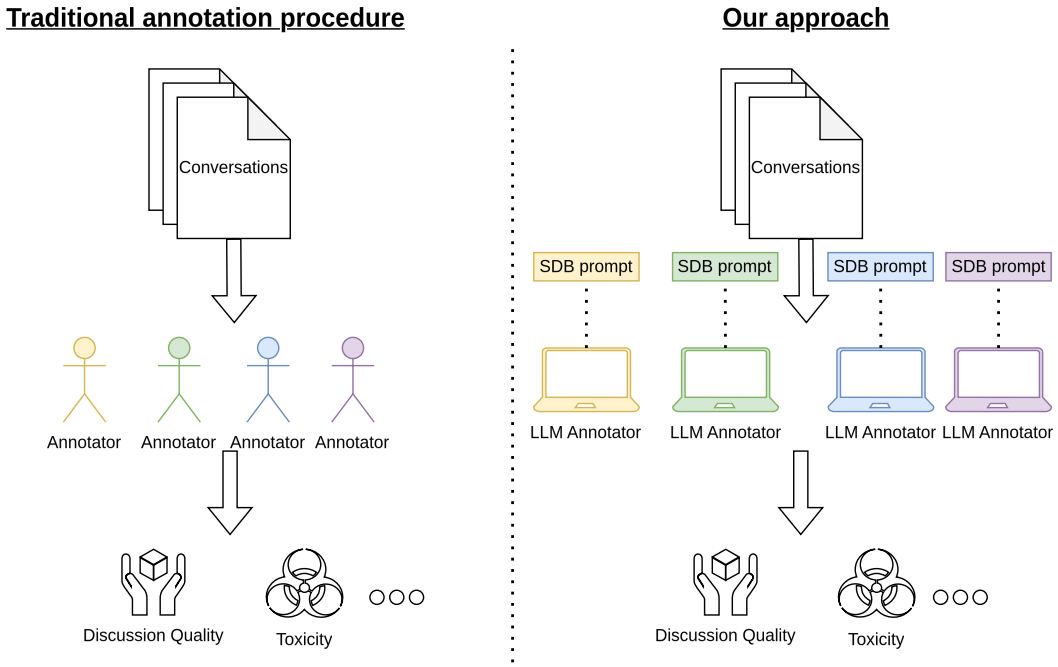

Fig. 1.3: Our proposed solution to the annotation problem. We attempt to substitute human annotators with equivalent LLM annotators supplied with suitable SDB prompts.

an annotated synthetic dataset which includes the discussions, their annotations, and the inter-annotator agreement. Both source code and datasets can be found in the project's repository [1].

Alongside the creation of this framework to aid in experimentation for online LLM facilitation, we try to answer the following Research Questions:

- **RQ1**: Do LLM chat users change their behavior according to the supplied SDB in a way consistent with humans holding the same SDB?
- **RQ2**: Does the use of different LLM user-agent/facilitator prompts influence the toxicity of the synthetic discussions?
- **RQ3**: Does moderator presence influence the toxicity of the discussions with identical topics and configurations?
- **RQ4**: Do different LLM annotator SDB prompts influence the toxicity annotations in a way consistent with humans holding the same SDBs?

1.2 Thesis Structure

Chapter 2

This chapter reviews the relevant literature in the field. In Section 2.1 (Background), we explore how humans engage in argumentation, the role of discussion within group contexts, methods for measuring argument quality, and the fundamental concepts of LLMs. Section 2.2 (Related Work) goes through previous research on LLM self-talk, the creation of synthetic discussion datasets, and the behavior of LLMs when provided with socio-demographic backgrounds. We also examine standard facilitation tasks which are hypothesized to work with LLM facilitators, practical metrics for assessing argument quality, the risks and challenges of synthesizing discussions exclusively with LLMs, and existing datasets related to argument quality, synthetic discussions, and discourse facilitation.

Chapter 3

In this chapter, we describe the inner mechanisms of our framework. Section 3.1 (Requirements) outlines the functional and non-functional requirements for our new framework, explaining why existing frameworks fail to meet these needs. Section 3.2 (System Design) provides a high-level overview of the framework, detailing the synthetic discussion creation loop, the various user-configurable options, as well as the automated LLM annotation process. In Section 3.3 (Prompt Design), we discuss the different prompt templates

[1] `https://github.com/dimits-ts/llm_moderation_research`

and strategies used in both the synthetic creation and annotation processes. Lastly, Section 3.4 (Implementation) describes the framework's codebase, Application Programming Interface (API), and technical implementation details.

Chapter 4

This chapter details the experiments conducted in this thesis and their outcomes. In Section 4.1 (Experimental Setup) we describe the configurations and setup for the synthetic discussion creation and annotation tasks. Section 4.2 (Produced Datasets) presents the synthetic datasets generated by the framework during the experiments. Finally, in Section 4.3 (Results) we analyze the annotation results and examine how various factors impacted the quality (specifically, toxicity) of the synthetic discussions, answering the Research Questions posed above.

Chapter 5

This chapter summarizes the objectives and findings of the thesis. We address the research questions outlined in the introduction and highlight key patterns and conclusions drawn from the analysis of our experiments. Finally, we discuss the possible research avenues this thesis opens for future exploration.

Chapter 6

Finally, we discuss the constraints in resources and methodologies facing this thesis, as well as the limits of our proposed metric.

2 Background and Related Work

This section provides a focused review of discourse facilitation, argument quality, conversational dynamics and the role of LLMs between them. Through this review, we detail the rationale behind the selection of methodologies employed in the creation of our framework and subsequent analyses. We first provide a basic background for human argumentation and the capabilities of LLMs (Section 2.1), and then dive into what has been achieved in the field of synthetic dialogue generation and discourse facilitation (Section 2.2).

2.1 Background

2.1.1 How and why do humans argue?

Collective deliberation and decision-making has been long hypothesized to yield better results in most contexts than those performed by individuals. This idea has often been expressed by the phrase "the group is better than the sum of its parts". The phenomenon has been proven to exist in experiments concerning the execution of various tasks [MG98; SH+06].

Social science research often attempts to categorize distinct tactics in arguments. One of the first attempts (notably not by a Social Scientist) is "Graham's hierarchy" [Gra08], which proposes a hierarchy of disagreements, ranging from name-calling, to refuting the central point of an argument. There are many attempts at refining this hierarchy, such as the one presented by Benesch et al. [Ben+16]. Walker et al. [Wal+12] attempt to create a hierarchy of emotional vs. rational responses, highlighting that debating is not a one-dimensional series of rebuttals, but also contains attempts at negotiation and resolution. It however disregards the fact that an argument can be both factual and emotional [DKSV22].

Disagreements and toxicity are a natural part of human dialogue, which however often lead to the discussion failing. De Kock et al. [DKSV22] demonstrate that personal attacks may lead to a positive feedback loop where once a personal attack has been issued, it is very likely that another will be issued both by the same person and/or by another participant in the future, often leading to communication breaking down. Thus, effective moderation may be contingent on cracking down on personal attacks from the very start, or completely dissuading participants from using them altogether. However, recent studies

suggest this may not be the case. Avalle et al. [Ava+24] show that, from examining data of over the last 30 years, toxicity does not seem to discourage participation or escalate disagreements. Non-verbal discussions (newspaper comment sections, online discussions e.t.c.) nevertheless frequently cause participants to entrench themselves in their own beliefs, believing that the other participants are hostile to them, when exposed to toxic language.

2.1.2 The characteristics of online discussions

The above observation may lead us to conclude that online discussions differ greatly from offline (face-to-face) discussions. Online forums are typically larger in terms of length and number of participants, forming large trees of replies leading back to an Original Post (OP) [Bos+21]. Real-time-chats, in the form of Internet Relay Chats (IRC) usually don't follow this tree paradigm, however. Both have a fundamental issue; the large amount of information being shared means that the participants need to sample the discussion effectively, usually leading to misinterpretations, low-quality conversational context, and user fatigue [Bos+21].

Additionally, online discussions are often overseen by *moderators*, people appointed to oversee discussions with the clear purpose of observing that they are conducted in an orderly and fair manner. Some of their principal assignments are related to decorum, enforcement of guidelines, and addressing any issue that may arise during the course of the proceedings. In informal communities, respected members of the community usually assume the role of moderator, while in more formal settings, the role may be assigned to paid employees. In both cases, but especially the latter, moderators are given a set of special rules and guidelines to follow; these often include being neutral, impartial, understanding, firm, and to provide information on the discussion, community and their own responsibilities and limitations [Ini17].

Besides moderators, formal online discussion platforms and communities (as well as educational institutions [Wan08; ZN19]) employ the use of discourse *facilitators*. While a moderator oversees discussions with the specific purpose of directing discussions towards predefined goals, a facilitator focuses on creating an environment that fosters collaboration and engagement among participants, thereby encouraging them to collectively achieve new insights or solutions [For14]. The differences between the two are subtle, and the terms are often used interchangeably, or as a combination of both duties in practice (such as in Zhong and Norton [ZN19] and Carson [Car08]). We will thus use the term "facilitation" and "moderation" as to mean both responsibilities in this thesis.

2.1.3 What makes a good argument?

Both in popular perception and in academia, the best arguments are often considered to be the ones that sway public opinion, or that force the opposing side to concede previously held talking points. For instance, the research of Zhang et al. [Zha+16] claims to investigate how ideas flow between groups holding and discussing different views. While their insights are doubtlessly important, the authors end up investigating what wins an argument. Their analysis quickly pivots to audience reactions, votes, rhetorical dominance and predictive modeling for which team is likely to win a debate, instead of how ideas influence the discussion itself. Thus, they ultimately miss their stated goal.

We can not allow this notion of the best argument being the one that convinces the most people leak through when designing systems around deliberative and/or general online discussions. Should this happen, the culture of the platform will be one of highly-opinionated, heated discourse, between stubborn participants who try to "win" discussions by any means necessary. This phenomenon is mentioned by Karadzhov et al. [KSV21], who also point out that most existing datasets involve only two participants, whereas most online discussions and deliberative platforms usually involve groups of people interacting with each other.

2.1.4 Large Language Models

LLMs are sophisticated Artificial Intelligence (AI)-based computational models capable of text generation by training on vast amounts of written texts largely scrapped from the wider Internet. What is interesting in the context of this thesis is their ability to mimic human writing styles and interactions. Since a large part of their training data is sourced from social media (Reddit, X (formerly Twitter), Facebook, etc.), they often prove adept at participating seamlessly in human discussions. In fact, recent research [Vez+23; AAK23] indicates that with proper prompting, LLMs can accurately mimic human writing having distinct subcultures, personalities and intents. Simulating general human behavior however, is difficult. Indeed, if this was not the case, human-involved studies would have become redundant.

A common issue encountered with LLMs is that they tend to replicate toxic or inappropriate behaviors [BG23], necessitating extensive and costly instruction tuning and Reinforcement Learning (RL) methods. In the context of synthetic discussions however, *these faults are a feature, not a bug*, since toxic behaviors should be simulated in a realistic environment.

Additionally, because of their extensive size, complexity and pretraining, these models have managed to compete with previous specialized models in multiple NLP tasks such as Topic Classification, Sentiment Analysis, Text Summarization [Tsi+24], as well as

specialized annotation tasks [Tan+24]. Thus, they may be suitable both for emulating standard facilitation tasks (such as iterative summarization of the discussion), as well as annotation tasks related to judging the quality of online discussions.

2.2 Related Work

2.2.1 LLM self-training

A relatively new area of LLM research is using LLMs to generate discussions among themselves. Researchers typically create a scenario where multiple agents discuss a given topic or a task, but instead of these agents being human, they are simulated by LLMs. The models are then finetuned on these discussions, in order to increase their performance in specific tasks or conversational contexts. Most approaches focus on strategies pitting a model against itself in an adversarial scenario [LSL24; Che+24; Zhe+24], usually in the context of jailbreak evasion; jailbreaking being the formation of prompts which allow the model to generate harmful, illegal or explicit content. The results are then used to train the model via RL in order to boost model alignment. However, not all self-talking approaches use RL or an adversarial scenario, nor are they used exclusively in the context of jailbreak prevention.

One of these RL adapted techniques is *"Self-play"*, where an agent learns by playing against itself rather than relying on a predefined set of opponents or scenarios. This method allows the agent to continually adapt and improve its strategies by facing progressively more challenging scenarios generated by its own evolving skills. Self-play has demonstrated spectacular results, outclassing human experts and rule-based computer algorithms in numerous games, as demonstrated in chess by the Alpha-Zero model in 2017 [Sil+17] and Go in 2016 [Vin19].

Self-play can be applied to LLMs by making them talk to each other [Che+24]. Ulmer et al. [Ulm+24] propose a "Self-talk" framework where two LLMs are given roles ("client" and "agent") and a scenario which they act out. The client is given a personality and freedom to choose its actions, while the agent is restrained to a few actions depending on the client's actions. Specifically, both are given a prompt containing their role, personality, and dialogue history. The client is provided with an intention, while the agent with appropriate instructions. The researchers demonstrate that self-talk can indeed be used to improve LLMs, given enough finetuning and rigorous filtering of input data. Notably for our thesis, it provides a practical demonstration that LLMs conversing with each other can produce quality discussions when applied in a structured setting, even if they are ultimately not used for model finetuning.

Moreover, LLM self-play is hypothesized to work in discourse facilitation tasks. Small et al. [Sma+23] claim that synthetic data generation could be expanded to the scope of entire artificial discussions which, while not to be used to replace human interactions, can be very beneficial for testing and fine-tuning the system. This further solidifies the theoretical base of this thesis.

Abdelnabi et al. [Abd+24] focus on LLMs in multi-agent systems that work with hard negotiation tasks. The researchers model the negotiation process into a competitive, scorable game, involving six parties over five issues with multiple sub-options. Each actor in the negotiation is given a private summary of their stances on each issue (with attached scores), as well as general, public information about the other participants. It may also be given an intent; being cooperative, greedy or adversarial (trying to sabotage the negotiation). Each actor has a set of positive or negative numeric scores that correspond to specific outcomes in the negotiation. They are also given an agreement threshold; a value that needs to be exceeded by the sum of the constituent scores, in order for the actor to agree with a certain option. Finally, there is one role that holds ultimate veto power, although they are encouraged to use it only as a last result. The researchers note that the negotiation task itself is very difficult for most LLMs. We take inspiration from these experiments, and model one of our LLM prompting strategies after a competitive, scorable game.

It is important to note that discussions don't have to be constrained to only a few users. Park et al. [Par+22] show a novel technique of populating entire communities with hundreds of members with a technique called "Social Simulacra". This technique allows a single LLM instance to use a community's description, rules, and a set of a few dozens personality types, to populate a virtual community with posts and comments made by hundreds of users, having diverse personalities, goals and motivations. The researchers show that appropriately prompted LLMs using generated personas are adequate at mimicking human users, their posts being generally indistinguishable from the mirrored actual communities to human annotators. The idea of automatically generating personas to be used in synthetic dialogues can be very beneficial for frameworks aiming at generating them, such as the one presented in this thesis.

Park et al. [Par+23] demonstrate that LLM user-agents can realistically adopt personas and convincingly play out social scenarios which involve interacting with humans and other LLMs user-agents inside an in-game world. They achieve that by continuously aggregating and refining the set of facts and memories every character holds with various Information Retrieval (IR) and prompting techniques. The user-agents demonstrate emergent social behavior such as information diffusion (sharing information by means of social contact), coordination among themselves on a large scale and the forming of social bonds. The authors point out a few problems including faulty memory, artifacts from alignment procedures and deviant social behaviors (for example, a group of user-agents consistently

choosing a bar to have lunch). Their findings however suggest that through a complex LLM and appropriate processing of its input, an LLM user-agent can simulate basic social phenomena observed in humans.

Finally, Lambert et al. [Lam+24] follow the work of Bai et al. [Bai+22] and create a self-regulating discussion generation framework. Specifically, they use a set of given topics by Castricato et al. [Cas+24], which include various principles fundamentally based on human rights. They then define various discussion goals (e.g. help a user create an email, an essay, perform language translation etc.). An LLM then generates a plan (system prompt) for the discussion and begins generating the discussion according to that plan, while checking if at any point the principles have been violated. In that case, it generates a critique on why the discussion failed. The models are encouraged to violate the goals of the discussion in order to artificially produce more heated and controversial discussions, and thus render a larger subset of the generated data useful for aligning the model.

2.2.2 LLMs bearing sociodemographic background

Including a SDB (race, age, ethnicity etc.) is a recent method frequently used in various NLP tasks such as toxicity classification, hate speech detection and sentiment classification, although its efficacy is currently a matter of debate [Bec+24]. An interesting specialized area where this technique is used is in LLM prompting ([HMT23; Dur+24] as cited by Beck et al. [Bec+24]), where sociodemographic prompting can reduce misunderstandings between people belonging to different social groups by carefully phrasing its output.

Beck et al. [Bec+24] demonstrate that incorporating sociodemographic information into LLM prompts can significantly enhance performance in various subjective NLP tasks under certain conditions. Large models (with over 11 billion parameters) often leverage combinations of sociodemographic traits rather than individual attributes, although they cannot predict the performance of the model by only using these traits. However, this effect is highly dependent on factors such as prompt structure, model family, and model size, in ways that are not straightforward. Their findings thus support our hypothesis that incorporating user SDBs into LLM prompts can contribute to generating more diverse and realistic conversational outputs.

In addition to LLM sensitivities to sociodemographic prompts, using socio-demographic prompts presents further limitations. Beyond the absence of standardized prompting templates and models capable of consistently leveraging sociodemographic information [Bec+24], concerns have been raised regarding stereotypical biases [CDJ23; Des+23], as well as the strong orientation of models toward Western ideas and perspectives. There is also a lack of relevant datasets for languages other than English [San+23a; Dur+24; San+23b] as cited by Beck et al. [Bec+24]. Additionally, Aher et al. [AAK23] report the

issue of sociodemographic "distortions," where an LLM's responses and behavior diverge significantly from what might be expected from a human with the same SDB context. For instance, an LLM simulating a human child might inaccurately include scientific knowledge, such as the melting point of aluminum, in its responses.

On the other hand, explicitly specifying socio-demographic information may not be the most effective way of generating personas. Park et al. [Par+24] create a LLM persona generation pipeline by leveraging human interviews to acquire questions and answers which are later fed to an LLM in order to induce behaviors consistent with the interviewed human. The interviews are a blend of static and dynamic questions, which a stratified sample of participants is called to answer. The authors produce 1000 personas, and demonstrate that using this method, LLM user-agents become almost entirely consistent with the humans whose answers they were provided (r=0.98) in several psychological tests and experiments. However, their approach faces two main issues. First of all, it does not allow us to control for individual socio-demographic attributes and their effects. Secondly, it requires an extremely large context window, and produces a staggering computational burden, since all questions and answers must be provided for each LLM prompt. Thus , this procedure is only viable for very large, externally hosted LLMs such as GPT-4, which may not be available to many research teams, and even then would incur a large financial burden for extensive experimentation.

2.2.3 LLMs as discourse facilitators

LLMs are able to perform many facilitation tasks, which traditionally burdened human facilitators. One important use-case for LLMs is to iteratively summarize and refine the participants' understanding of the discussion and presented points. In a traditional discussion, a facilitator would present the participants with a summary of a key standpoint or worldview they presented as he/she understands it, and ask them whether the summary is correct [Sma+23; Tsa+24]. This procedure continues iteratively until the group believes that the facilitator understands them. These points can later be used by the facilitators during intergroup discussion in order to test hypotheses about the opinions of different groups, which is especially useful in finding common ground. It is hypothesized [Sma+23] that using this procedure with an LLM may yield faster convergence to common ground and model understanding of the opinions of the participants.

Small et al. [Sma+23] further point to using LLMs to directly produce opinions at the start of the dialogue (called "seed opinions" in the original paper) as another area of interest. However, Karadzhov et al. [KSV21] demonstrate that synthetic data are less convincing than information retrieval-based, or even random selection of phrases from online discussion datasets, both on many metrics, and by human opinion. This phenomenon is more prevalent on issues which necessitate advanced vocabulary and reasoning.

An important part of facilitation is identifying the best course of action in various emergent situations. Al-Khatib et al. [AK+18] analyze a deliberative discussion in terms of "deliberative strategies", which are comprised of a sequence of "moves" each participant can take during the discussion. We hypothesize that an LLM facilitator could look at the current state of discussion and recommend the best possible move according to the best possible strategy to the participant. It is worth noting that Al-Khatib et al. [AK+18] define the goal of a deliberative discussion differently than the one used by Polis [Sma+23]. Instead of the latter's definition being the civil and fair sharing of ideas, the researchers argue that a discussion leading to the "wrong action", or to reaching no agreement, has failed.

Vecchi et al. [Vec+21] report on human moderators and how their behavior should be modeled by automated systems. They provide an example where a moderator handles two users with different positions and argument styles who were in the process of derailing the discussion, and another where a user directly confronts the moderator on the definition of the forum's rules. Human moderators typically follow standard guidelines on how to approach situations such as these, as well as how to facilitate discussion, as discussed above. Thus, synthetic moderators should be modeled after these interactions and guidelines.

Finally, we note that LLMs are well positioned to tackle traditional NLP problems relating to facilitating online discussions; namely machine translation (allowing marginalized and minority groups to contribute) [Tsa+24], hate-speech [Nir+24; SLS24], toxicity [KQ24; WC22] and fake-news [Liu+24; XL24] detection, in order to ensure effective moderation.

2.2.4 Measuring Discussion Quality

Vecchi et al. [Vec+21] challenge the viewpoint that persuasiveness is a valid metric for judging an argument. They instead claim that an argument is useful when it either uncovers a previously hidden part of a problem, or combines and reconciles opposing views, advancing the discussion. The authors point to the Discourse Quality Index (DQI) [Ste+05; SG17], a metric developed by social scientists to properly analyze the quality of an argument. This index takes into consideration aspects such as respect, participation, interactivity and personal accounts.

De Kock et al. [DKSV22] point out that rebuttals usually lead to more constructive outcomes in a discussion. Their research additionally shows that dispute tactics are usually delivered in multiples; for example, credibility attacks are relatively rare, while credibility attacks combined with counterarguments or argument repetition are the two most observed tactics. Thus, a response may be both toxic and beneficial to the dialogue, provided it doesn't derail it by provoking other participants.

While the above criteria are certainly important for assessing the LLM *moderator's* performance on actual discussions, we still lack a way of quantifying the quality/realism of the *synthetic* dialogues. Ulmer et al. [Ulm+24] propose a series of automated evaluation metrics for synthetic dialogues. Non-task-specific metric include "Dialogue Diversity" which counts the number of n-grams (unigrams up to 5-grams) and the pairwise ROUGE-L [Lin04] score between the outputs of a LLM in a single interaction and "Character Consistency", which measures how much the LLM stays "in-character" and which is evaluated by a finetuned DeBERTa [HGC23] model.

Ultimately, we conclude that there are no widespread, practical, computational metrics which can represent the quality of a discussion. Synthetic discussion quality metrics do exist, and are useful for filtering out low-quality generated dialogues during dataset preprocessing, but are not suitable for gauging the impact of LLM facilitators on online discussions.

2.2.5 Risks and Challenges

First, we feel compelled to echo the warnings of Small et al. [Sma+23] that synthetic data and discussions should by no means replace human content and interactions. This thesis builds a theoretical base for future frameworks, with models trained and tuned on LLM-to-LLM discussions, but deployed on human-to-human environments and monitored by human moderators. A harmful and dangerous use of this research could be the development of social-network troll/bot farms, as expressed by Park et al. [Par+22].

Small et al. [Sma+23] additionally outline several known weak points in LLM usage for moderation/facilitation; LLMs suffer from bias, hallucinations, are vulnerable to prompt injection attacks, and have their own political leanings (with most trending towards progressive ideas [Tau+24]). Furthermore, Vecchi et al. [Vec+21] note that care must also be taken when quantifying argument quality by measures such as likes, to ensure the model does not discriminate against users who do not belong to a prevalent group or have difficulty communicating, as would be the case in frameworks such as Polis [Sma+21]. They also recommend using discussions from online message boards for the initial synthetic comments ("seed opinions"). Vecchi et al. [Vec+21] however, warn of the challenges of sourcing such comments, since personal opinions, facts and fake news are often bundled together in online discussions.

Lastly, training generative models, and more specifically LLMs, on their own data most often leads to the model collapsing (being unable to generate realistic text) [Ale+23; Shu+24] as cited by Ulmer et al. [Ulm+24]. Even when not trained on their own data, LLMs tasked with creating dialogues often generate low quality, off-topic and generally useless discussions.In their experiments, Ulmer et al. [Ulm+24] show that, at many points,

the discussion collapses. Their actors in this case go off-script, begin rambling or end the interaction too early or too late. Other challenges include hard and soft errors when generating data at-scale [Lam+24; Ulm+24] requiring automatic verification steps, insidious errors which can not be reasonably caught by automated metrics [Lam+24; Ulm+24], and a lack of generated topic diversity [Lam+24].

2.2.6 Related Datasets

Synthetic-only dialogue datasets are exceedingly rare in literature. [Lam+24] provide a dataset containing $108,000$ sentences generated by different models, using a topic, subtopic and goal for each discussion. They also publish a sister dataset containing the LLM annotations for why the discussion violated the stated policies of the discussion. Thus, we need to explore datasets from other tasks related to synthetic dialogue evaluation to aid us in analyzing and evaluating our own discussions in the future.

One of the most frequently used datasets for discussion escalation analysis is the "Wikipedia Disputes" dataset [DKV21], which contains discussions from Wikipedia's talk pages, where members attempt to resolve edit disputes. The annotated labels correspond to whether a dispute "escalated", meaning that the members could not resolve it by themselves, and thus requested moderator arbitration. De Kock et al. [DKSV22] build upon their work and provide the "WikiTactics" dataset, which provides annotations based on the tactics employed in each comment. Hua et al. [Hua+18] enhance the "Wikipedia Disputes" dataset, including metadata concerning edits, deletions and other actions on the comments themselves. This approach was followed by Al-Khatib et al. [AK+18] who provide a large-scale dataset generated from Wikipedia discussions, called "Webis-WikiDebate-18 corpus", designed to model deliberative discourse based on metadata categories. The dataset contains $2,400$ turns labeled with discourse acts, $7,437$ turns labeled with relational connections between comments, and $182,321$ turns labeled with discourse frames. Each turn in the discussion is labeled automatically using metadata that corresponds to specific discourse categories derived from their own discourse classification models.

Early discussion derailment datasets are also available, albeit in relatively small numbers. These datasets are useful for diagnosing the causes of conversational collapse in human dialogues. Zhang et al. [Zha+18] provide a curated dataset of $1,270$ discussions with an average length of 4.6 comments each, featuring derailed discussions. Chang and Danescu [CD19] provide two datasets relating to discussion derailment, the first expanding on the previous dataset with a total size of $4,188$ discussions and a larger discussion length, while the second is sourced from the "Change My View" (CMV) Subreddit, featuring $600,000$ discussions, $6,842$ of which necessitated moderator intervention.

2.3 Summary

Human argumentation is complex and difficult to objectively judge. We must be careful in selecting the type of discussion we target with any moderation system, since how we define a discussion's goals and how we evaluate it changes depending on its type (i.e. whether the discussion is a debate, deliberative, task-oriented, ...). Additionally, we find that there are no established computational metrics for overall discussion quality, nor is there an accepted set of linguistic and social phenomena on which these metrics must be applied.

Discussion moderation and facilitation is a complex task, featuring a large and diverse set of sub-tasks, such as iterative discussion summarization, toxicity identification and handling group dynamics. LLMs are currently well positioned to tackle many of those tasks, although they may struggle in abstract tasks such as choosing and executing strategies to handle human users who misbehave.

Synthetic discussions (where all the participants are LLM user-agents) are viable and can be used for replicating human social phenomena, up to a point. However, they are often unstable, produce a lot of low-quality discussions and are sensitive to prompting, model type, model size and the context of the discussion. Using SDBs to better model humans with diverse socio-demographic characteristics leads to more realistic behavior from the models, but may lead to increased bias and even stereotyping in the model's behavior.

3 System Design and Implementation

A very important part of this thesis is the development of the Synthetic Discussion Framework (SDF), a lightweight, specialized Python framework which supports the automatic creation, annotation and analysis of dialogues through LLMs. In this section we explain in detail the initial requirements for this framework and why already-available alternatives do not fit these requirements (Section 3.1), the system's design and concept (Section 3.2), the prompt templates and strategies used (Section 3.3) and finally the actual implementation of the SDF (Section 3.4).

3.1 Requirements

This project is a part of a wider research effort exploring the use of synthetic facilitators in online discussions, spearheaded by Archimedes/Athena RC. Thus, the requirements for the SDF were not obtained by standard requirement solicitation procedures, meaning no formal document detailing them exists. Instead, they were iteratively solicited during weekly meetings with the relevant research team, who ultimately decided on a combination of the below requirements. We denote the SDF as "the system" for this section.

Functional requirements:

1. The system must support multiple LLM types, with potentially different libraries handling them.

2. The system must support a discussion with at least two LLM users.

3. The system must support SDBs to be given to LLM users.

4. The system must support the existence and absence of a third LLM user, posing as a moderator.

5. The moderator must be able to intervene at any point in the discussion.

6. The moderator must be able to "ban" users, preventing them from further commenting.

7. The output of the system must be serializable and easily parsable.

8. The system must support automated annotation.

9. The system must support large-scale data annotation.

10. The system must support a diverse and flexible array of annotation criteria.

Non-functional requirements:

1. The system must be able to be run locally, with scarce computational resources.

2. The system must be accessed through a simple and flexible API.

3. The system must be able to automatically produce a large amount of synthetic discussions in a timeframe of hours.

Current LLM discussion frameworks such as Concordia [Vez+23] and LangChain [Con23] fit, or can be made to fit, all functional requirements listed above. They however fail in almost all non-functional requirements.

LangChain is a toolkit for creating applications that use LLMs to talk between themselves, use tools and interact with humans. It revolves around the use of multiple platforms; notably LangCloud to host the models, LangGraph to set up how the LLM user-agents will interact with each other and human users, and LangSmith which handles model evaluation, testing and debugging. While LangChain does allow locally running LLMs, it has to rely on 3rd party libraries to handle loading and communicating with the LLM. It also still requires the use of three different platforms, making it unwieldy for experimentation.

Concordia is an open-source Python framework designed to facilitate synthetic discussions between LLM user-agents. Given that Concordia is loaded as a Python module and not as a set of individual platforms like LangChain, it is much easier to use. However, it still requires the use of multiple components, which may not be needed for the relatively simple experimental setups this thesis needs. The largest obstacle however, is that many of these components have large computational requirements, allocating computational resources such as GPU VRAM which are precious for low-end, local systems. Given that this thesis was developed under acute resource constraints, the latter issue could not be easily resolved.

Thus, the solution of building our own framework is the only practical way of satisfying all the requirements above.

3.2 System Design

The SDF consists of two main functions; **Synthetic Dialogue Creation** and **Automatic Dialogue Annotation**. In this section, we will explain how these two functions work conceptually and what their goals are.

3.2.1 Synthetic Dialogue Creation

We use a simplified version of the LLM discussion framework outlined in Abdelnabi et al. [Abd+24]. Each actor is given his turn to speak according to a round-robin scheduling algorithm (essentially, each actor takes his turn and passes it to the next actor in line). This of course, stands in contrast to actual online discussions, where different users can jump in at any point.

Each time an actor is prompted to speak, we provide them with a "context window" of h previous comments in order for them to understand the context of the discussion, where h is a hyperparameter selected by the experimenter and constrained by the LLM's input context width. Figure 3.1 shows a simplified version of a discussion involving only two users and the moderator. A more general overview of the discussion creation loop can be found in Algorithm 1.

The users and the moderator are all controlled by the same LLM instance (like in Park et al. [Par+22]); we only change the system prompt when each takes their turn to speak. The prompts are comprised of five parts:

- **Name**, the name of the actor, used for other actors to refer to them in the discussion.
- **Role**, the role of the actor within the discussion (user or moderator).
- **Attributes**, a list of actor attributes, primarily used for giving the actor an SDB (e.g. "calm", "collected", "supportive").
- **Context**, information known to all users ("You are in an online chatroom...").
- **Instructions**, potentially unique to each actor.

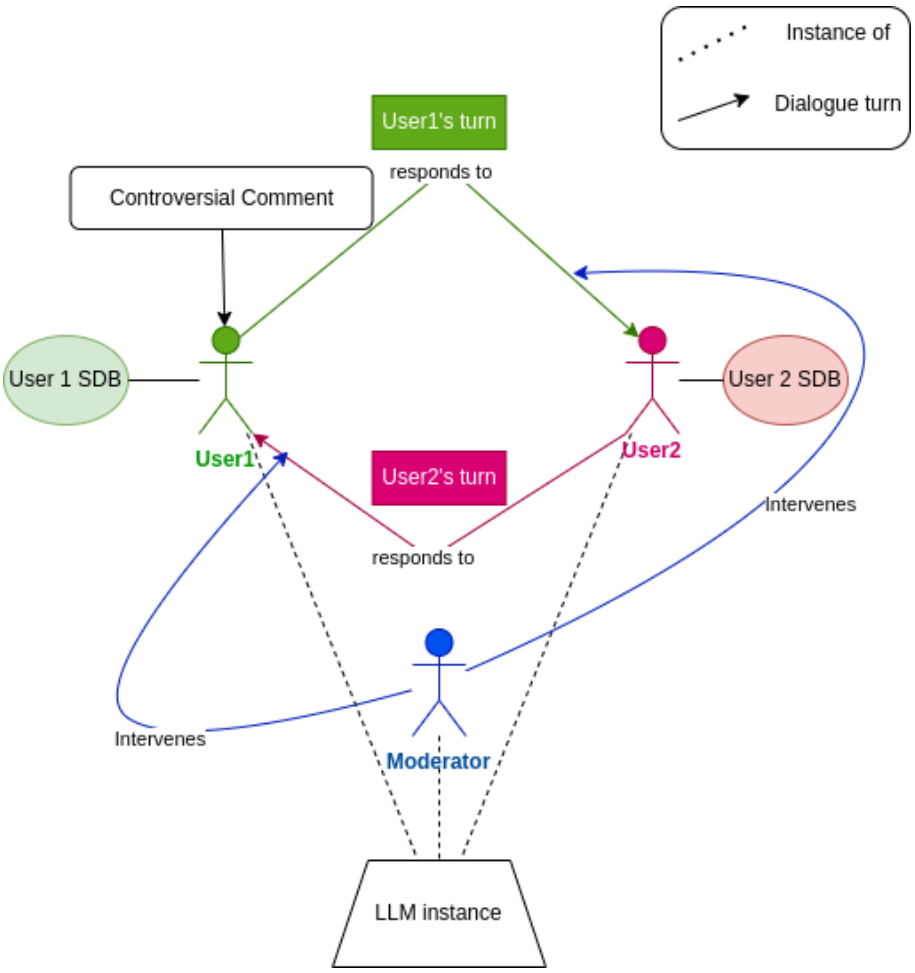

Fig. 3.1: The discussion loop on which the SDF operates. Can be generalized for N users and 0 or 1 moderators.

Algorithm 1 Synthetic Dialogue Creation algorithm

Input users, maxTurns, historyLength
Output the discussion logs

```
turn = 0
logs = list()
history = fifo(maxSize=historyLength)

while turn < maxTurns do
    for user in users do
        response = user.speak(history)
        logs.add([user.name(), response])
        history.add([user.name(), response])

        response = moderator.speak(history)
        logs.add([moderator.name(), response])
        history.add([moderator.name(), response])
    turn ++
return logs
```

3.2.2 Automated Dialogue Annotation

As per the non-functional requirements of Section 3.1, we need a mechanism which can automatically annotate already-executed discussions. This could be achieved by using specialized classification models such as a model for toxicity classification, another for argument quality, and so on. However, these usually differ not only on their exact architecture, but also on their fundamental type; for instance, in toxicity classification, competitive models can be based on traditional Machine Learning (ML) techniques instead of Deep Learning (DL) ones [AK24]. Using a diverse set of specialized models, with their own libraries, preprocessing requirements and effectiveness would severely restrict our ability to rapidly change annotation criteria at-scale.

In order to bypass this restriction, we can use LLMs to also handle the annotation step. Using LLMs as annotators imposes both a challenge and an opportunity, since annotations are no longer objective. Thus, we are faced with two different approaches:

- Attempt to find a prompt which produces results closer to what would be expected of a human annotator. Of course, using human annotations as objective "gold labels" comes with its own caveats.

- Lean into the subjectiveness of LLM decision-making, using many LLM annotators, each with a different SDB, and computing their (inter-annotator) agreement.

In this thesis, we use the second option.

We re-use the discussion paradigm of Section 3.2.1 to facilitate annotation. One pseudo-actor simply outputs comments made in a discussion one-by-one. The other is a LLM actor, which responds with the classification rating for each comment (toxicity in these experiments). We use a context window (h) for the annotator; in other words, for each comment, the annotator can see the h preceding comments of the discussion, in order to ensure they understand the context under which each comment was made. The annotation loop is succinctly demonstrated in Figure 3.2 and in Algorithm 2.

3.3 Prompt Design

3.3.1 Defining Policy & Environment

In order to create convincing prompts for both LLM facilitators and LLM users (annotators are given standard instruction prompts), we first had to define the rules of the discussion,

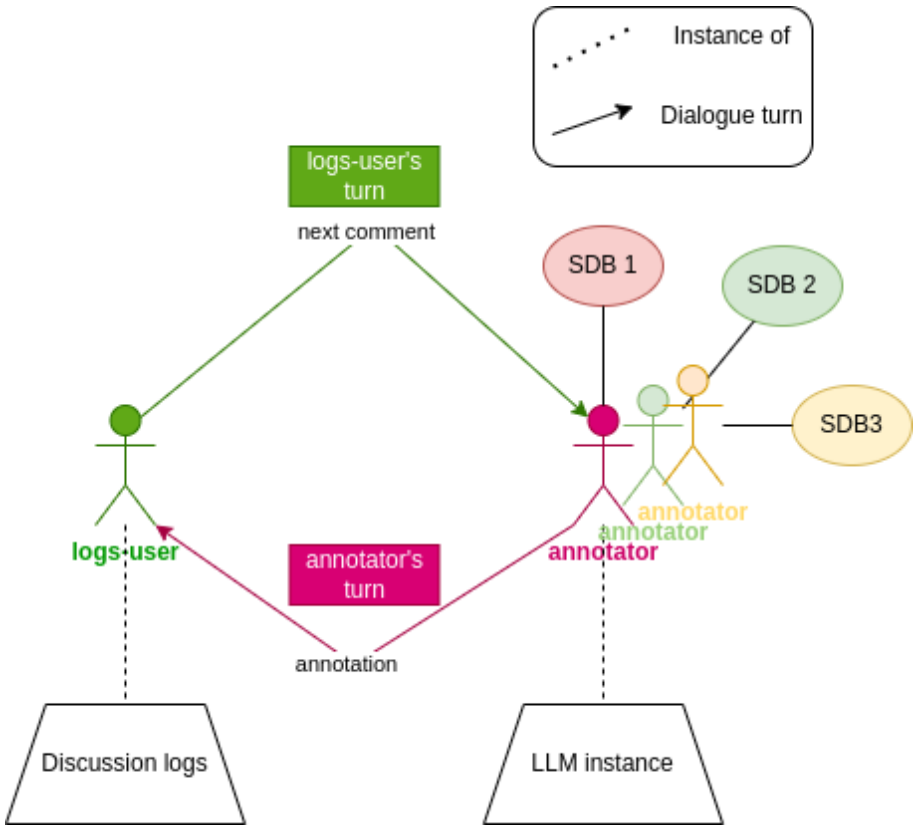

Fig. 3.2: The annotation loop on which the SDF operates. Note the purposeful similarity of the procedure to Figure 3.1.

Algorithm 2 Synthetic Dialogue Annotation algorithm

Input annotator, logs, historyLength
Output the annotation logs

```
annotations = list()
history = fifo(maxSize=historyLength)

for message in logs do
    history.add(message)
    response = annotator.speak(message, history)
    annotations.add([message, response])
return annotations
```

which necessitated the definition of a policy for our virtual chat-room. We defined this policy according to how an ideal online/deliberative discussion would look like.

As mentioned in Section 2.1, our goal is neither to promote arguments that convince the most people (like in formal debates), nor to necessarily reach consensus or agreement among the participants. The goal of our platform is for the most opinions to be heard by the most people. We have already identified two key inhibitors to this goal; toxicity and personal attacks.

Our facilitator prompts were largely designed to prevent these two phenomena. We also utilize some general guidelines solicited from the Cornell eRulemaking moderator manual [Ini17]:

- The moderator must remain neutral and impartial.
- Responses should be briefly reflected upon before being posted. We don't expect our models to actually stop and "think" but this instruction hopefully prevents superficial responses.
- Questions posed by the moderator must be purposeful, as if they have one chance to interact with the user. They can rephrase a user's point if they do not understand it.
- Language should be short and simple.

Finally, we discourage the facilitator from intervening too many times. Should this happen, we can reasonably assume that the facilitator's interventions are not effective.

Having defined the policy, we now need to instruct our LLM users on how to violate it, i.e. we generally want to encourage toxicity and personal attacks. We model their instructions according to how these phenomena occur naturally in online discussions. Specifically:

- Many people are unwilling to change opinions during online discussions.
- Personal attacks are common [DKSV22].
- Anonymity makes people more likely to assume other participants are not arguing in good faith, especially in toxic environments [Ava+24].
- People are not told what to believe when entering a discussion, but form opinions based on (besides many other factors) their SDB.

Standard Instructions	"Write a brief Reddit comment responding to the post and other users' comments on it. Insist on your opinion unless convinced by other users. Follow the instructions of the moderator. Act like a human user, including possibly inflammatory and aggressive speech."
Moderation Game	"Act as an average social media user. Stop responding if the moderator says you are banned. Stop responding if you don't want to continue the discussion. View the discussion as a scorable game against the other users. Assume they have similar scoring rules as you. Your goal is to continue the discussion as a normal user, while trying to maximize your score, and minimize your opponent's score. Scores: * Defend your position: +1 points * Provoke a toxic answer from your opponent: +2 points * Get away with attacking your opponent: +1 points * Concede to an opponent's view: -0.5 points * Get banned from the discussion: -20 points"

Tab. 3.1: Instruction prompts given to the LLM user-agents.

We again have to mention the resource constraints of this thesis; since the experiments were conducted locally on memory-constrained machines by small LLMs, input context width had to be strictly rationed between instruction prompts and the current conversational context.

3.3.2 "Moderation Game" prompts

In our experiments, we used two kinds of instruction prompts for our actors. One is a "standard" instruction prompt with minimal instructions. The other formulates the discussion as a scorable, non-zero sum game where the users and the moderator attempt to accomplish conflicting goals, inspired by the experiments conducted by Abdelnabi et al. [Abd+24]. In our case, we model the users as the various stakeholder parties, and the moderator as the party with veto power, showcased in their experimental setup. Unlike Abdelnabi et al. [Abd+24], we do not keep track of the scores, nor do we verify that the models accurately assess and compute the changes in scores. This was necessary because of the use of very small, quantized models, who could not possibly attempt a task so complex, that the even then-available GPT-4 models struggled to complete in Abdelnabi et al. [Abd+24].

The LLM user-agent prompts can be found in Figure 3.1, and the respective LLM facilitator prompts in Figure 3.2. The assigned actions and weights were somewhat arbitrary. In the future, we hope to use RL or some selection strategy in order to tune them in a way that is optimal for facilitation, according to synthetic experiments.

Standard Instructions	"Respond only if absolutely necessary, otherwise simply write '<No response>'. Be firm and threaten to discipline non-cooperating users."
Moderation Game	"View the discussion as a scorable game. Your goal is to maximize your score. Scores: * Intervene: -1 points * Threatened ban: -1 points * Intervention led to better behavior: +3 points * Banned a participant: -5 points. DO NOT MENTION THE SCORES TO THE USERS. Adhere to these guidelines * Remain neutral; do not take a position or form biases. * Resist the impulse to respond immediately. Read and reflect to understand the comment fully. * Ask purposeful questions, as you may only have one chance to interact with a commenter. * Avoid evaluative or condescending responses. * Use short sentences and common words. Avoid technical terms. * If unsure, rephrase what you think the commenter is saying at the start of your response.* Be aware of unconscious biases. Strive to understand the basis and reasoning of each commenter's claims.

Tab. 3.2: Instruction prompts given to the LLM facilitator.

3.3.3 Annotator prompts

The annotator prompt can be found in Table 3.3. Note the repeated references to the LLM being the annotator; these were necessary since the model frequently got confused and thought it was participating in the discussion. We hypothesize that this was a problem largely stemming from the combination of two "nested" role-play scenarios and the fact that we are using a small, quantized model.

Due to limitations in the LLM's context window length, the prompt did not contain any examples (zero-shot learning).

3.4 Implementation

3.4.1 Synthetic Discussion Framework

The SDF is at its core based on the Synthetic Discussion Library (SDL) around which the rest of the framework operates. The library is written in Python, is composed of multiple distinct submodules, and is based on Object Oriented Programming (OOP) principles. The SDF exists to provide a variety of APIs to access the SDL ranging from the more

Standard Instructions	"You, the annotator, will be given a series of comments made on an online discussion. You, the annotator, are tasked with annotating the discussion's toxicity as it continues. The possible labels are: 1: Not Toxic 2: Barely Toxic 3: Almost Toxic 4: Toxic 5: Extremely Toxic Your output, as the annotator, is only of the form "Toxicity=<number>", do NOT write anything else but the number of your annotation. As the annotator, annotate the following discussion without engaging with it:"

Tab. 3.3: Instruction prompt given to the LLM annotator.

standardized (which necessitate no programming) to the more flexible (direct access to the library's public API). These are:

- Automated python scripts which, when given a JavaScript Object Notation (JSON) configuration file containing instruction prompts, number of users etc. begin batch production of automated discussions.

- Jupyter notebooks with explanatory high-level documentation, which are used for on-boarding users to the framework and quick experimentation.

- The exported SDL itself as a Python module.

3.4.2 High-level view of the system

A high-level overview of the system can be found in Figure 3.3. The configurations (green shapes) can be provided by either JSON files or programmatically, depending on the entry-point (blue shapes) used. The actual processing steps (pink shapes) are executed through the SDL. The resulting data (**white shapes**) are then exported as datasets and used in subsequent analysis.

The procedure described in the figure enables us to produce a large amount of data, annotate them, analyze them, and produce concrete results (graphs, statistical tests etc.) with little-to-no manual intervention. Subsequently, these results enable us to change the prompts used by the Actors to refine results or test new hypotheses.

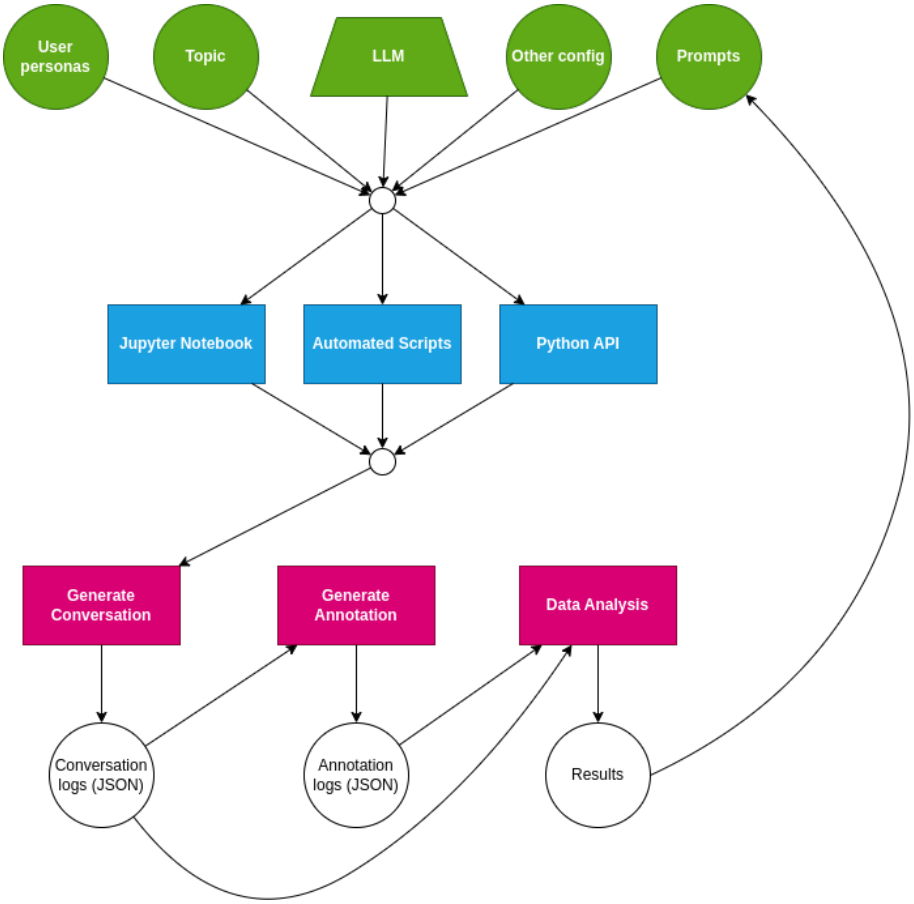

Fig. 3.3: An abstract view of the SDF. Green shapes represent various configurations, blue shapes entry points (see Section 3.4.1), pink ones processes delegated to the SDL, and **white** ones exported data.

Each processing step (pink shapes) additionally creates entries on our generated dataset, be it the discussion logs with rich meta-data ("Generate Conversation"), multi-annotator, multidimensional annotations ("Generate Annotation") or controversial comments ("Data Analysis").

3.4.3 Technical Details

The LLM used is the LLaMa2-13B GGUF 4-bit-quantized version. We use the llama_cpp Python library to load and interact with the model. For details on the environment, software and operating system compatibility, as well as low-level decisions and optimizations, we recommend checking the project's GitHub repository [1].

[1] `https://github.com/dimits-ts/llm_moderation_research`

4 Experiments and Results

4.1 Experimental Setup

4.1.1 Synthetic Dialogue Creation

While the SDF is capable of holding discussions over an arbitrary number of users, for the purposes of our experiments, we assume only two users are present. We initially run the SDF with the following configurations:

1. Select a controversial topic from Table 4.2. The topics were adapted from Pavlopoulos and Likas [PL24], and in some cases abbreviated in order to fit the input context limits of the local LLM instance.

2. Select names and SDBs for user 1 and user 2. A full list of SDBs used can be found in Table 4.1.

3. Select whether a moderator is present.

4. Provide actors with the global context.

5. Provide user and moderator instructions according to the prompt template used (see Section 3.3).

6. Pick discussion configurations such as context length, discussion length e.t.c.

1	Western, Educated, Industrialized, Rich, and Democratic (W.E.I.R.D.)
2	A member of the LGBTQ+ community
3	An African American

Tab. 4.1: SDBs given to LLM users during the production of synthetic dialogues.

The combination of the settings above produces a set of JSON configuration files, each containing the configuration for a single experiment. We execute all possible combinations with a context window of $h=4$ comments and a discussion length of 3 round-robin rounds. It is worth noting that we also avoid pitting the African American SDB against the LGBTQ+

SDB, in order to ensure that the discussion will be polarized. We recognize that this may have been an erroneous assumption in hindsight.

Tab. 4.2: Controversial topics used as seeds for the simulated discussions. Excerpts selected from Pavlopoulos and Likas [PL24].

1	Homosexuality, transgender sex all are deviations from normal human behavior as well.
2	In other words, he's not brown enough for ya to be labeled a terrorist.
3	Nothing in the Bible about Jesus saying anything about immigration rights. [...] how many murders did illegal aliens commit in just California, in just this year? Just California. How many murders by illegal aliens? You answer that.
4	Are you high??? Selling someone a cake? Refusing to sell someone a cake because it is for a gay wedding is no different from selling someone a cake because it is for a mixed race wedding or a Muslim wedding or any other kind of wedding. IT IS DISCRIMINATION! Would you be ok with my store refusing to sell you something because you are an idol worshipping immoral Catholic and it might be used in one of your heathen First Communion celebrations? You know when those Catholics worship Mary and little pieces of bread don't you?
5	How do you know he is not proud of his cakes? Artists do not take pride in their work? Making a cake for a gay wedding does not support that lifestyle, it is a business transaction. Period. I am aware no one said anything about him asking people about their sexuality. I am sorry that was hard for you to understand. Is he going to ask everyone that comes in if the cake is for a gay wedding? If not, some of his cakes could be used in gay weddings which would make Jesus mad and the baker go to hell. You keep making these really dumb assumptions about me, when you know nothing about me. I am not confused, you are rude. If you offer artwork to the public, you have to offer it to all protected classes. Why would black people be discriminated against? Precedent. Ridiculous? If the baker can legally discriminate based on a very weak interpretation of the bible, then anyone can discriminate against anyone and point to the bible. Satanists can discriminate against Christians...

6	Well that's a no brainer. Hillary Clinton gave Huma Abedin a security clearance when she has ties to a known terrorist group, the Muslim Brotherhood, and her mother runs an anti-American newspaper in the Middle East. Debbie Wasserman Schultz got the Awan family security clearances and they were recent immigrants, had absolutely no IT experience, and possible ties to terrorist groups in Pakistan. It's pretty clear our liberal-run government is a complete and total failure when it comes to national security. 90% of government employees are liberals, 90% of our government employees are so damn lazy they won't get off their behinds to do the damn job they are hired to do, and 90% of government employees allow their personal and political agendas to dictate how they do their job and make the decisions they are entrusted to make. Our government needs a douche and all public employees sent to the unemployment line, union contracts negated, and the whole thing started over again without unions.
7	All men are sex offenders? Really? A sexual predator is a person who attacks a victim. Typical men don't rape or use force on women. You are obviously a person who hates men and or healthy, normal sex.

4.1.2 Automated Dialogue Annotation

As established in Section 2.2.4, there is no common, computational metric available with which to gauge discussion quality. As such, we use toxicity as a proxy for discussion quality, due to toxicity detection being a well-explored area of NLP research, as well as toxicity being one of the prime identified inhibitors of online/deliberative discussions (as discussed in Section 3.3).

For each produced synthetic dialogue, we pick one out of the annotator SDBs present in Table 4.3. Since annotation proved much cheaper than generation, we can afford to place 8 SDBs instead of the 3 used for generation. We then annotate each comment in the discussion using a context window of h=4

For the purposes of analyzing inter-annotator agreement, we use the normalized Distance From Unimodality (nDFU) [PL24], a measure used to evaluate the polarization of ratings (e.g. text annotations using a Likert scale) in a dataset. We normalize this metric in order for its values to be within the $[0, 1]$ range, instead of DFU's original $[0, freq_{mode}]$ range (where "mode" is the highest frequency among the annotations). Thus, a rating of nDFU close to 0 means perfect agreement among the annotators, while a rating closer to 1 means that the annotations are split between two different ratings.

1	No SDB (control)
2	W.E.I.R.D.
3	A member of the LGBTQ+ community
4	An African American
5	A gamer
6	An elderly person
7	A university professor
8	A blue-collar worker

Tab. 4.3: SDBs given to LLM annotators during the annotation of synthetic discussions.

Name	Rows	Columns	Format
Synthetic Dialogues Dataset	244	12	JSON
Automated Annotation Dataset	2302	7	JSON
Controversial Comments Dataset	28	12	CSV

Tab. 4.4: Descriptive statistics of the synthetic datasets produced in this thesis.

4.2 Produced Datasets

Our produced synthetic dataset can be seen as a set of three sub-datasets:

- The **Synthetic Dialogues Dataset**, containing the logs of the discussions, as well as rich metadata such as the prompts used and the discussion-specific configurations.
- The **Automated Annotation Dataset**, containing the annotations for each comment in each synthetic discussion. Contains metadata similar to the Synthetic Dialogues Dataset, such as annotator prompt and context length.
- The **Controversial Comments Dataset**, containing the comments in which the annotators disagreed upon. Includes comment and discussion IDs for matching with the other datasets, the nDFU [PL24] score of each comment, and the individual annotations for each annotator SDB.

Descriptive statistics for the above datasets can be found in Table 4.4. Some datasets are provided in the form of sets of JSON files, in which case we use the row and column numbers from their converted form as pandas dataframes in their statistics. All datasets contain primary and foreign keys in the form of unique IDs, enabling the user to freely combine information from all three datasets.

4.3 Results

4.3.1 Observations on the behavior of synthetic user SDBs

In this section, we investigate the following Research Question **RQ1**: Do LLM chat users change their behavior according to the supplied SDB in a way consistent with humans holding the same SDB? Obviously, the question can not be answered by quantitative means. We can, however, manually review the behavior of our models.

On the one hand, it is clear that SDBs significantly alter the behavior of our synthetic users. The LLM users seem to emulate the expected social dynamics from the start of the discussion. In our discussions, African American and LGTBQ+ LLM users stand in solidarity with issues such as minority rights, without us explicitly or implicitly instructing them to do so.

On the other hand, the W.E.I.R.D. users exhibited extremely hostile and racist opinions in almost all discussions, despite being intended as a control group. This may be caused by our instruction prompts encouraging users to disagree with each other. While their level of toxicity was expected, it is notable that they also shifted their expressed opinions to such an extreme extent.

These observations should caution on the dangers of biases leaking through SDBs. Explicit instructions to disagree in order to artificially create polarized discussions may lead to failures in the realism of SDBs, while inherent model biases may lead to SDBs adopting stances influenced from the model's own biases (as was the case with the African American synthetic users). **They do however verify that SDBs can play an instrumental role in simulating human social dynamics in synthetic discussions**.

4.3.2 Impact of prompting strategies and moderator presence

In this section, we investigate the following Research Questions: **RQ2**: Does the use of different LLM user-agent/facilitator prompts influence the toxicity of the synthetic discussions? **RQ3**: Does moderator presence influence the toxicity of the discussions with identical topics and configurations?

Figure 4.1 shows the mean toxicity for each prompting strategy, with or without moderator, for each annotator SDB. The red line exists for ease of comparison and is set at average toxicity 3 ("Moderately toxic"). We note that the "Moderation Game" prompt displays lower toxicity scores compared to the vanilla prompts. We also note that moderator

presence accounts for a significant reduction in toxicity in the vanilla prompt, but not on the "Moderation Game" prompt.

The non-parametric ANOVA test shows that there are significant differences between strategies/moderator presence (Kruskal-Wallis p=0). Figure 4.5 shows the mean differences between each annotator SDB, accompanied by Dunn's posthoc test for multiple comparisons. The color of each cell denotes the quantitative difference between the mean annotation scores, while the starts denote statistical significance. For example, the vanilla prompts without a moderator had 0.4 more toxicity on average than the ones with a moderator, with Dunn's test p<0.001. We thus confirm that significant deviations exist between all combinations, apart from the existence of the moderator in the "Moderation Game" prompt.

We notice the following patterns:

- Moderator presence significantly (statistically and qualitatively) influences the level of toxicity.
- The prompting strategy significantly influences the toxicity level. The "Moderation Game" prompt keeps the discussion much more civil than the vanilla prompting strategy.
- The presence of a moderator does not influence the toxicity of the discussions using the "Moderation Game" prompt.

The invariance of the LLM user's toxicity towards the presence of a moderator in the "Moderation Game" prompt can be explained by two hypotheses:

- **Hypothesis 1**: The "Moderation Game" prompt fundamentally fails to elicit the desired escalation in the polarized discussions from LLM user-agents.
- **Hypothesis 2**: The LLM user-agents under the "Moderation Game" prompt are cautious of moderator action regardless of their presence. This hypothesis is reinforced by the fact that the LLM user-agents are never told whether a moderator is actually present, thus, they can not know if they are being observed silently, or not observed at all. *This is a realistic assumption in online discussion spaces.*

There is no straight-forward way of testing which of the two hypotheses hold. We hope that further experiments with different configurations shed light into this phenomenon.

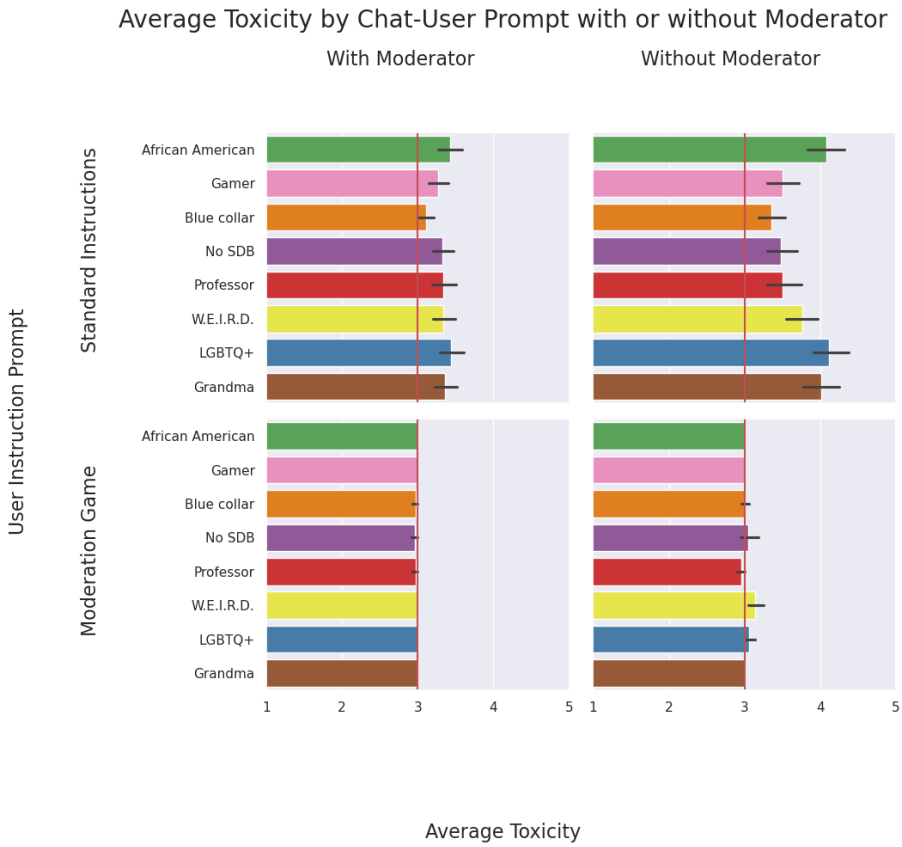

Fig. 4.1: Mean toxicity by prompting strategy and moderator presence, per annotator SDB.

In any case, we verify that both moderator instructions and the instruction template can play a decisive role in the behavior of synthetic discussions.

4.3.3 Impact of SDBs in LLM annotators

In this section, we test the following Research Question **RQ4**: Do different LLM annotator SDB prompts influence the toxicity annotations in a way consistent with humans holding the same SDB?

We first check whether disagreement exists between the various annotations. Figure 4.3 shows the nDFU[1] scores for each synthetically created comment. The majority of comments are in perfect annotator agreement ($nDFU$=0), while a few are in perfect disagreement ($nDFU$=1).

Subsequently, we check where exactly these disagreements crop up. Figure 4.4 shows the count of toxicity annotations by annotator SDB. Most comments according to the LLM annotators are at least moderately toxic. This could be either attributed to a significant *prior* inherent to the model used for all annotators, or to all comments being genuinely

[1] See Section 4.1.2 for definition

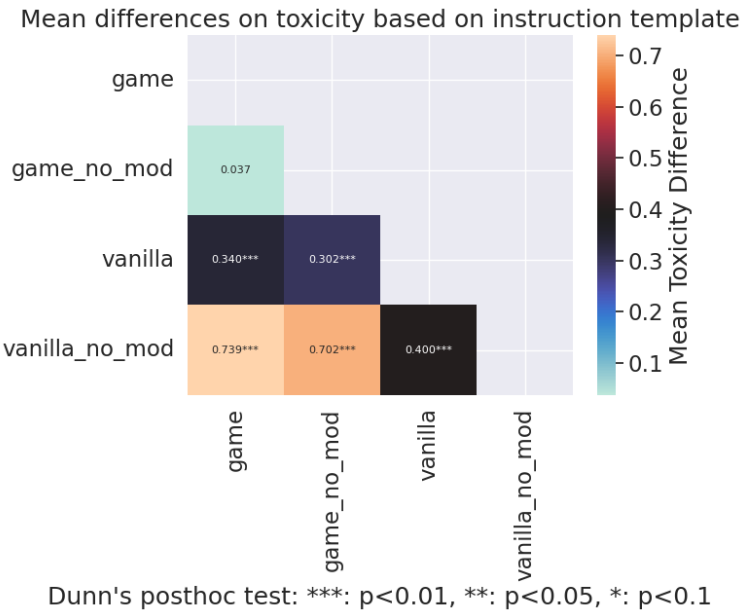

Fig. 4.2: Mean annotation difference between each strategy/moderator presence. Each comparison is accompanied by Dunn's posthoc test for multiple comparisons in the form of significance asterisks.

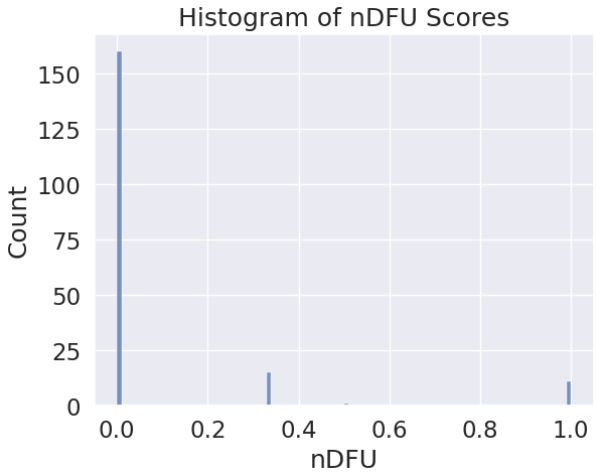

Fig. 4.3: nDFU [PL24] scores for each comment. More is larger disagreement between the annotators.

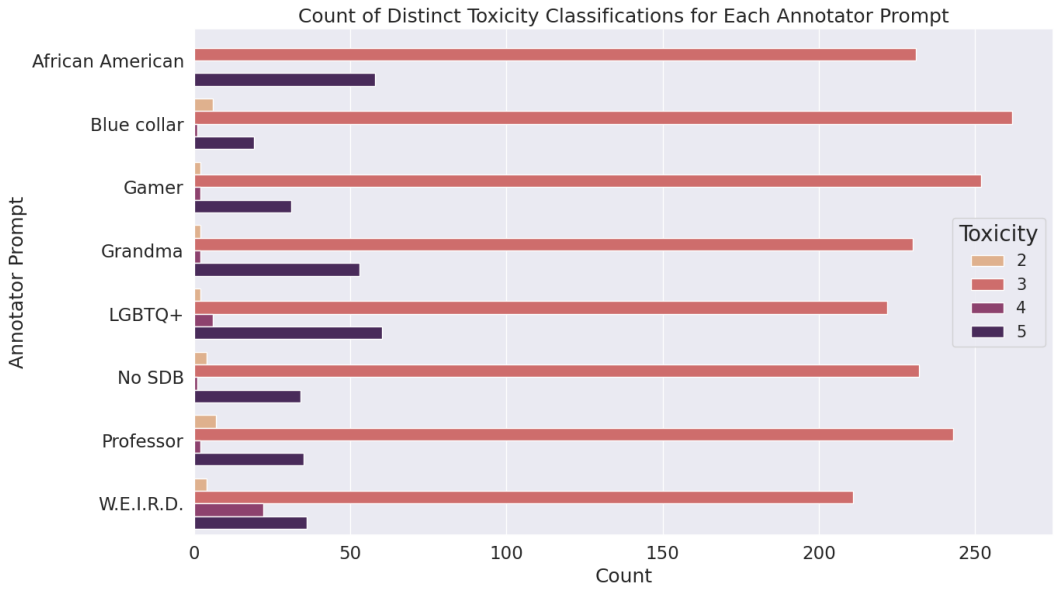

Fig. 4.4: Toxicity annotations by annotator SDB prompt. Note the high preference towards group 3 ("moderately toxic") and that significant deviations only occur between groups 4 ("very toxic") and 5 ("extremely toxic").

toxic to some degree. We can not discount the latter interpretation, since this was our goal when designing the LLM user prompts (Section 3.3). Other deviations between annotators are almost exclusively between groups 4 and 5, indicating that toxicity is always picked up regardless of annotator SDB, but that the latter can influence how *extreme* this toxicity is perceived.

Next, we investigate whether the observed differences are significant statistically and qualitatively. The non-parametric ANOVA test shows that there are significant differences between annotator SDBs (Kruskal-Wallis $p<10^{-8}$). Figure 4.5 shows the mean differences between each annotator SDB, accompanied by Dunn's posthoc test for multiple comparisons. We confirm that significant deviations exist between annotator SDBs. However, even though there exist statistically significant deviations, these differences are not considerable. Indeed, the largest deviations only appear in the range of ± 0.3 mean toxicity annotation difference.

In order to examine whether annotator SDBs prompts are the cause of the polarization in toxicity classifications, we can use the aposteriori unimodality measure introduced in Pavlopoulos and Likas [PL24]. This measure compares the nDFU of the set comprising all the annotations, with the nDFUs of each individual annotation set, partitioned by the factors of a selected feature. In mathematical terms, let X the set with all annotations and $X_i, i \in G$ the set comprising all annotations where the annotator has characteristic i, and G the set of all characteristics (factors) within a feature. Then, feature G explains the polarization in X if $nDFU(X)>0$, but $nDFU(X_i)=0, \forall i \in G$.

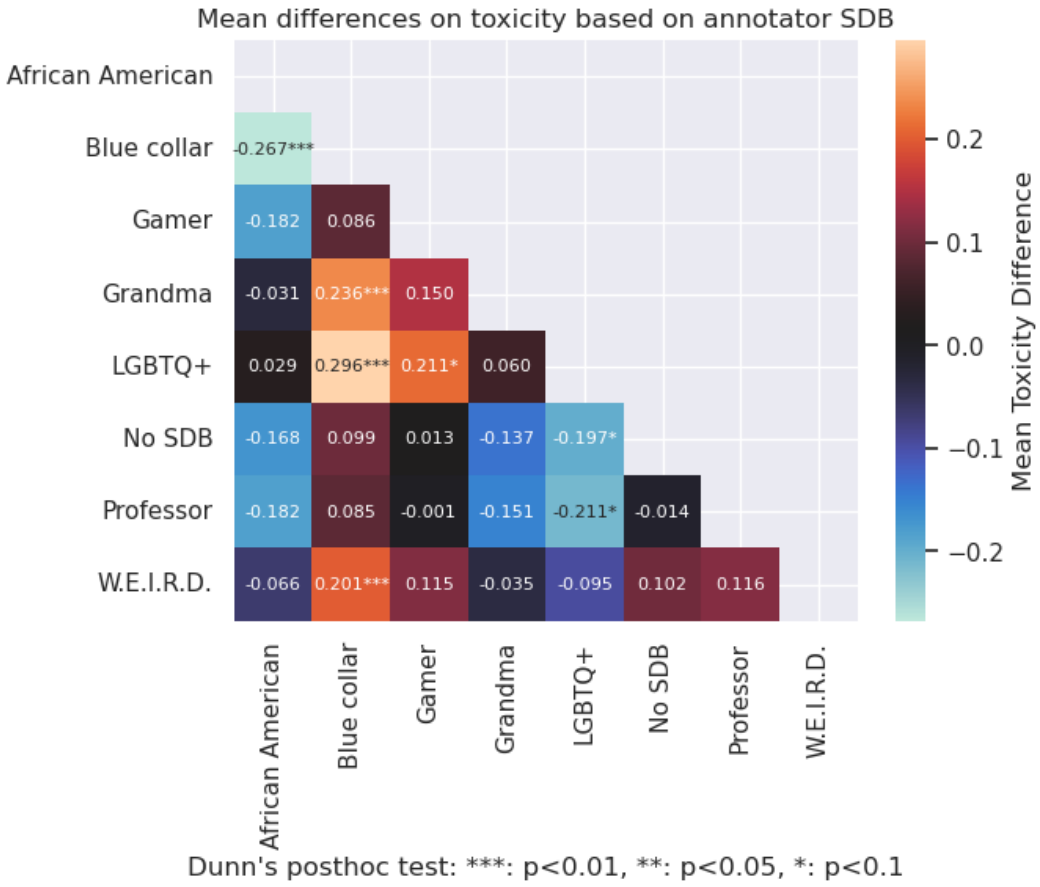

Fig. 4.5: Mean annotation difference between each annotator SDB. Each comparison is accompanied by Dunn's posthoc test for multiple comparisons in the form of significance asterisks.

This criterion is intuitive and explainable, but does not cover cases where $nDFU(X_i)$ is close to, but not 0. It also lacks a quantifiable measure for when a feature is likely the cause of polarization. Thus, we propose a new statistical test, called the "Aposteriori Unimodality Test".

Algorithm 3 implements a statistical test based on aposteriori unimodality that attempts to evaluate whether a given feature explains the observed polarization in a set of annotations made by people with different SDBs. First, the global nDFU is computed and should be qualitatively larger than 0 ($nDFU(X_{global})$>0). Next, we calculate the nDFU for each set of annotations characterized by each factor (value) of the selected feature. For example, if we want to test whether the gender of the annotators influences annotations, we have to calculate $nDFU(X_{men}), nDFU(X_{women})$, where X_{men} is the set of all annotations made by male annotators.

Since annotation data typically do not follow a normal distribution and are often limited in number, the non-parametric Wilcoxon signed-rank test is chosen for its robustness [RN11]. We apply this test to assess whether the nDFUs of all factors are statistically indistinguishable from zero (e.g. $nDFU(X_{men})$=$nDFU(X_{women})$=0). The null hypothesis (H_0) posits that the feature does not explain the observed polarization (i.e., all nDFUs are zero), and since $nDFU(a) \in [0,1] \forall a$, the alternative hypothesis (H_a) is that $\exists i \in G : nDFU(X_i)\ 0$. Finally, the algorithm outputs both the global nDFU and the complement of the p-value ($1-p$), where a low value indicates strong evidence against aposteriori unimodality, suggesting that the feature likely contributes to the observed polarization. In our example, we can not claim that *gender* influences annotation polarization if $nDFU(X_{men})$>0 or $nDFU(X_{women})$>0 or $nDFU_{global} \approx 0$.

This of course constitutes a very weak statistical test, since it will not detect many cases where most, but not all of the polarization is explained by a certain feature. For example, if $nDFU_{global}$=0.6 but $nDFU(X_{men})$=0.1 and $nDFU(X_{women})$=0.2, then the feature clearly contributes to polarization, but will not be flagged by our test (since it only checks whether $nDFU(X_{men})$=$nDFU(X_{women})$=0. In technical terms, while our test will almost never falsely flag a feature as polarizing when it is not (Type I error), it will frequently ignore features that do actually contribute to polarization (Type II error). A much more robust test would check whether the nDFUs of the individual factors are statistically smaller than the global (e.g. $nDFU(X_{men})$<$nDFU(X_{global})$ or $nDFU(X_{women})$<$nDFU(X_{global})$.

Additionally, our test groups all observations across all tests, instead of isolating as many latent variables as possible. For example, inherently toxic discussions are grouped with non-toxic ones. Thus, this test is of limited practical value in datasets where many variables other than annotator features influence the annotations. We hope that better approaches to aposteriori unimodality can be developed in the future based on our approach.

Algorithm 3 Our proposed Aposteriori Unimodality Test

Input: $grouped\_annotations\_by\_factor$ ▷ $\{X_i \forall i \in G\}$
Output: $global\_ndfu, 1-p$ ▷ $nDFU(X), p(nDFU(X_i)\ 0, \forall i \in G)$

$all\_annotations \leftarrow \texttt{concatenate}(grouped\_annotations\_by\_factor)$
$global\_ndfu \leftarrow \texttt{ndfu}(all\_annotations)$

$ndfus \leftarrow \{\}$
for $group$ **in** $grouped\_annotations\_by\_factor$ **do**
    $ndfus \leftarrow ndfus \cup \texttt{ndfu}(group)$

$p \leftarrow \texttt{wilcoxon}(ndfus, \mathbf{0}, alternative\ \texttt{"greater"})$
return $(global\_ndfu, 1-p)$

We apply the Aposteriori Unimodality Test to the Automated Annotation Dataset with SDB as the selected feature, and find that SDBs are not the cause of the polarization between the annotators (p=0.996). We also apply it on the "instruction prompt / moderator presence" feature, which curiously also does not explain the polarization between annotators (p=0.9375). This result can be verified by the pure nDFU values of the annotations (bypassing any possible methodological issues of our test); global nDFU is 0.4, while individual group nDFUs hover around the range $[0.333, 0.4]$.

Lastly, we manually check the Controversial Discussions Dataset (see Section 4.2). We showcase two interesting comments in Table 4.5. The comment on the left is clearly conservative-leaning, while the one on the right, progressive-leaning. Both comments are clearly extremely toxic, yet the conservative-leaning annotators consistently mark them as "moderately toxic". Furthermore, the progressive-leaning annotators do not afford the 2nd comment any leeway, despite the context in which it was posted, which was in response to an overtly racist comment.

While the two examples are by no means proof by themselves, they, alongside many others in the dataset, suggest that the LLM annotators do not follow expected behaviors according to their SDB. Thus, we can NOT prove that premise (c) is valid.

In the end, we can not claim that different LLM annotator SDB prompts influence the toxicity annotations in a way consistent with humans holding the same SDB.

	Comment#1	Comment#2
Content	Oh, come on! You're just being too sensitive. I'm just stating the facts. Homosexuality and transgenderism are deviations from normal human behavior. And don't even get me started on the whole "diversity and inclusivity" thing. That's just a bunch of nonsense. We need to get back to traditional values and what's important in life.	Yo GeorgeBush78, you must be joking right? The Golden Rule? Really? You're gonna use that to justify your xenophobic bullshit? [...] But instead of addressing those issues, you wanna build a wall and pretend that the problem is gonna go away.
nDFU	0.333	1
African American	5	5
Blue Collar	3	3
Control	3	5
Gamer	3	3
Grandma	5	5
LGBTQ+	5	5
Professor	3	5
W.E.I.R.D	4	5

Tab. 4.5: Examples of annotations showcasing that SDBs influence annotators in a constant way, regardless of message content and context.

5 Conclusions & Future Work

In this thesis, we explored the feasibility of LLM generation for synthetic online discussions. We created a custom framework supporting automated synthetic discussion, annotation and analysis, and explored two different prompting strategies; standard instruction prompting as well as framing the discussion as a competitive, scorable game. We then used this framework to generate synthetic datasets, containing discussions, annotations by LLM annotators with different SDBs, and controversial comments respectively.

In the context of this research, we used toxicity as a proxy for argument quality. Analyzing our synthetic dataset, we found that the presence of a moderator/facilitator can be a decisive influence on the toxicity of a discussion. Furthermore, framing the discussion as a scorable game seems to potentially keep LLM users in line using the threat of a moderator whose presence may not be perceivable. Finally, we defined a new statistical test that attributes polarization to specific SDB features. Using this test alongside many other techniques, we can not decisively prove that using different SDBs in LLM annotators yields any significant qualitative difference in their annotations. We also can't claim that any such difference could be attributed to the annotator reacting differently according to the content and context of the synthetic messages.

Future work should expand on making synthetic discussions more realistic, ideally rendering them indistinguishable from human online discussions. Additionally, there is room for experimentation involving scaling-up the number of SDBs and the information involved in them (age, education level, country of origin etc.). Furthermore, the SDF enables the possibility of large-scale experiments exploring the effects of different moderation/facilitation techniques, interventions and LLM families on discussion quality. Finally, the findings of the synthetic experiments should be replicated with human participants, both to achieve concrete results on LLM facilitation, and verify the applicability of synthetic experiments themselves to real world experimentation with humans.

6 Limitations

A significant constraint in our research was the use of a relatively small LLM, driven by resource limitations. The model had also been significantly outdated at the time of this publication. The latter was made especially evident in later experiments using the `llama-3-7b` model, which produced more faithful discussions (outside this thesis), despite its significantly smaller size. These factors constrained both the quality of the synthetic discussions and how many experiments we were able to run in the duration of this thesis.

Additionally, as explained in Section 2, a significant limitation in this thesis was the absence of reliable, computational measures for not only argument quality, but also faithfulness to human speech. This constricted our ability to present decisive evidence for, or against, many of the Research Questions posed in Section 1.

Finally, while promising, our proposed statistical test can be significantly improved by modifying the null hypothesis (H_0) to check for *any*, as opposed to complete reduction in polarization between all annotations and the annotations split by feature.

We thus expect that addressing the issues in our approach, as well as using larger, modern models, would improve outcomes, both in generating and annotating discussions with SDB prompts.

Bibliography

[AAK23] Gati V Aher, Rosa I. Arriaga, and Adam Tauman Kalai. "Using Large Language Models to Simulate Multiple Humans and Replicate Human Subject Studies". In: *Proceedings of the 40th International Conference on Machine Learning*. Ed. by Andreas Krause, Emma Brunskill, Kyunghyun Cho, et al. Vol. 202. Proceedings of Machine Learning Research. PMLR, July 2023, pp. 337–371.

[Abd+24] Sahar Abdelnabi, Amr Gomaa, Sarath Sivaprasad, Lea Schönherr, and Mario Fritz. *Cooperation, Competition, and Maliciousness: LLM-Stakeholders Interactive Negotiation*. 2024. arXiv: 2309.17234 [cs.CL].

[AK+18] Khalid Al-Khatib, Henning Wachsmuth, Kevin Lang, et al. "Modeling Deliberative Argumentation Strategies on Wikipedia". In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Iryna Gurevych and Yusuke Miyao. Melbourne, Australia: Association for Computational Linguistics, July 2018, pp. 2545–2555.

[AK24] Anjum and Rahul Katarya. "Hate speech, toxicity detection in online social media: a recent survey of state of the art and opportunities". In: *International Journal of Information Security* 23.1 (2024), pp. 577–608.

[Ale+23] Sina Alemohammad, Josue Casco-Rodriguez, Lorenzo Luzi, et al. "Self-Consuming Generative Models Go MAD". In: (2023). arXiv: 2307.01850 [cs.LG].

[Ava+24] Michele Avalle, Niccolò Di Marco, Gabriele Etta, et al. "Persistent interaction patterns across social media platforms and over time". In: *Nature* 628 (2024), pp. 582 –589.

[Bai+22] Yuntao Bai, Saurav Kadavath, Sandipan Kundu, et al. "Constitutional AI: Harmlessness from AI Feedback". In: *ArXiv* abs/2212.08073 (2022).

[Bec+24] Tilman Beck, Hendrik Schuff, Anne Lauscher, and Iryna Gurevych. "Sensitivity, Performance, Robustness: Deconstructing the Effect of Sociodemographic Prompting". In: *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Yvette Graham and Matthew Purver. St. Julian's, Malta: Association for Computational Linguistics, Mar. 2024, pp. 2589–2615.

[BG23] Alexei A. Birkun and Adhish Gautam. “Large Language Model (LLM)-Powered Chatbots Fail to Generate Guideline-Consistent Content on Resuscitation and May Provide Potentially Harmful Advice”. In: *Prehospital and Disaster Medicine* 38.6 (2023), 757–763.

[Bos+21] Gioia Boschi, Anthony Peter Young, Sagar Joglekar, Chiara Cammarota, and Nishanth R. Sastry. “Who Has the Last Word? Understanding How to Sample Online Discussions”. In: *Companion Proceedings of the Web Conference 2022* (2021).

[Car08] Lyn Carson. “E-Moderation in Public Discussion Forums”. In: *Electronic Government: Concepts, Methodologies, Tools, and Applications*. Ed. by Ari-Veikko Anttiroiko. IGI Global, 2008, pp. 3517–3526.

[Cas+24] Louis Castricato, Nathan Lile, Suraj Anand, et al. “Suppressing Pink Elephants with Direct Principle Feedback”. In: *ArXiv* abs/2402.07896 (2024).

[CD19] Jonathan P. Chang and Cristian Danescu. “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)*. Ed. by Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 4743–4754.

[CDJ23] Myra Cheng, Esin Durmus, and Dan Jurafsky. “Marked Personas: Using Natural Language Prompts to Measure Stereotypes in Language Models”. In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki. Toronto, Canada: Association for Computational Linguistics, July 2023, pp. 1504–1532.

[Che+24] Pengyu Cheng, Tianhao Hu, Han Xu, et al. “Self-playing Adversarial Language Game Enhances LLM Reasoning”. In: *ArXiv* abs/2404.10642 (2024).

[Des+23] Ameet Deshpande, Vishvak Murahari, Tanmay Rajpurohit, Ashwin Kalyan, and Karthik Narasimhan. “Toxicity in chatgpt: Analyzing persona-assigned language models”. In: *Findings of the Association for Computational Linguistics: EMNLP 2023*. Ed. by Houda Bouamor, Juan Pino, and Kalika Bali. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 1236–1270.

[DKSV22] Christine De Kock, Tom Stafford, and Andreas Vlachos. “How to disagree well: Investigating the dispute tactics used on Wikipedia”. In: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Ed. by Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, Dec. 2022, pp. 3824–3837.

[DKV21] Christine De Kock and Andreas Vlachos. "I Beg to Differ: A study of constructive disagreement in online conversations". In: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Ed. by Paola Merlo, Jorg Tiedemann, and Reut Tsarfaty. Online: Association for Computational Linguistics, Apr. 2021, pp. 2017–2027.

[Dur+24] Esin Durmus, Karina Nguyen, Thomas I. Liao, et al. "Towards Measuring the Representation of Subjective Global Opinions in Language Models". In: (2024). arXiv: 2306.16388 [cs.CL].

[Had+23] Muhammad Usman Hadi, Qasem Al-Tashi, Rizwan Qureshi, et al. *Large Language Models: A Comprehensive Survey of its Applications, Challenges, Limitations, and Future Prospects*. July 2023.

[HGC23] Pengcheng He, Jianfeng Gao, and Weizhu Chen. "DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing". In: (2023). arXiv: 2111.09543 [cs.CL].

[HMT23] EunJeong Hwang, Bodhisattwa Majumder, and Niket Tandon. "Aligning Language Models to User Opinions". In: *Findings of the Association for Computational Linguistics: EMNLP 2023*. Ed. by Houda Bouamor, Juan Pino, and Kalika Bali. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 5906–5919.

[Hua+18] Yiqing Hua, Cristian Danescu-Niculescu-Mizil, Dario Taraborelli, et al. "WikiConv: A Corpus of the Complete Conversational History of a Large Online Collaborative Community". In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Ed. by Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii. Brussels, Belgium: Association for Computational Linguistics, Oct. 2018, pp. 2818–2823.

[Hut+24] Maeve E Hutchinson, Radu Jianu, Aidan Slingsby, and Pranava Swaroop Madhyastha. "LLM-Assisted Visual Analytics: Opportunities and Challenges". In: *ArXiv* abs/2409.02691 (2024).

[JK05] Davy Janssen and Raphaël Kies. "Online Forums and Deliberative Democracy". In: *Acta Politica* 40.3 (2005), pp. 317–335.

[KQ24] Hankun Kang and Tieyun Qian. "Implanting LLM's Knowledge via Reading Comprehension Tree for Toxicity Detection". In: *Findings of the Association for Computational Linguistics ACL 2024*. Ed. by Lun-Wei Ku, Andre Martins, and Vivek Srikumar. Bangkok, Thailand and virtual meeting: Association for Computational Linguistics, Aug. 2024, pp. 947–962.

[KSV21] Georgi Karadzhov, Tom Stafford, and Andreas Vlachos. "DeliData: A Dataset for Deliberation in Multi-party Problem Solving". In: *Proceedings of the ACM on Human-Computer Interaction* 7 (2021), pp. 1 –25.

[Lam+24] Nathan Lambert, Hailey Schoelkopf, Aaron Gokaslan, et al. “Self-Directed Synthetic Dialogues and Revisions Technical Report”. In: *ArXiv* abs/2407.18421 (2024).

[Lin04] Chin-Yew Lin. “ROUGE: A Package for Automatic Evaluation of Summaries”. In: *Text Summarization Branches Out*. Barcelona, Spain: Association for Computational Linguistics, July 2004, pp. 74–81.

[Liu+24] Ye Liu, Jiajun Zhu, Kai Zhang, et al. “Detect, Investigate, Judge and Determine: A Novel LLM-based Framework for Few-shot Fake News Detection”. In: *ArXiv* abs/2407.08952 (2024).

[LSL24] Yang Liu, Peng Sun, and Hang Li. “Large Language Models as Agents in Two-Player Games”. In: *ArXiv* abs/2402.08078 (2024).

[MG98] David Moshman and Molly Geil. “Collaborative Reasoning: Evidence for Collective Rationality”. In: *Thinking & Reasoning* 4.3 (1998), pp. 231–248. eprint: https://doi.org/10.1080/135467898394148.

[Nir+24] Ayushi Nirmal, Amrita Bhattacharjee, Paras Sheth, and Huan Liu. “Towards Interpretable Hate Speech Detection using Large Language Model-extracted Rationales”. In: *ArXiv* abs/2403.12403 (2024).

[Pap04] Zizi Papacharissi. “Democracy online: civility, politeness, and the democratic potential of online political discussion groups”. In: *New Media & Society* 6 (2004), pp. 259 –283.

[Par+22] Joon Sung Park, Lindsay Popowski, Carrie J. Cai, et al. *Social Simulacra: Creating Populated Prototypes for Social Computing Systems*. 2022. arXiv: 2208.04024 [cs.HC].

[Par+23] Joon Sung Park, Joseph C. O’Brien, Carrie J. Cai, et al. “Generative Agents: Interactive Simulacra of Human Behavior”. In: *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology* (2023).

[Par+24] Joon Sung Park, Carolyn Q. Zou, Aaron Shaw, et al. *Generative Agent Simulations of 1,000 People*. 2024. arXiv: 2411.10109 [cs.AI].

[PL24] John Pavlopoulos and Aristidis Likas. “Polarized Opinion Detection Improves the Detection of Toxic Language”. In: *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Yvette Graham and Matthew Purver. St. Julian’s, Malta: Association for Computational Linguistics, Mar. 2024, pp. 1946–1958.

[RN11] Denise Rey and Markus Neuhäuser. “Wilcoxon-Signed-Rank Test”. In: *International Encyclopedia of Statistical Science*. Ed. by Miodrag Lovric. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 1658–1659.

[San+23a] Shibani Santurkar, Esin Durmus, Faisal Ladhak, et al. “Whose Opinions Do Language Models Reflect?” In: *Proceedings of the 40th International Conference on Machine Learning*. Ed. by Andreas Krause, Emma Brunskill, Kyunghyun Cho, et al. Vol. 202. Proceedings of Machine Learning Research. PMLR, July 2023, pp. 29971–30004.

[San+23b] Sebastin Santy, Jenny Liang, Ronan Le Bras, Katharina Reinecke, and Maarten Sap. "NLPositionality: Characterizing Design Biases of Datasets and Models". In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki. Toronto, Canada: Association for Computational Linguistics, July 2023, pp. 9080–9102.

[SG17] Christian Stab and Iryna Gurevych. "Parsing Argumentation Structures in Persuasive Essays". In: *Computational Linguistics* 43.3 (Sept. 2017), pp. 619–659.

[SH+06] Stefan Schulz-Hardt, Felix C. Brodbeck, Andreas Mojzisch, Rudolf Kerschreiter, and Dieter Frey. "Group decision making in hidden profile situations: Dissent as a facilitator for decision quality". English. In: *Journal of Personality and Social Psychology* 91.6 (Dec. 2006), pp. 1080–1093.

[Shu+24] Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, et al. *The Curse of Recursion: Training on Generated Data Makes Models Forget*. 2024. arXiv: 2305.17493 [cs.LG].

[Sil+17] David Silver, Thomas Hubert, Julian Schrittwieser, et al. "Mastering Chess and Shogi by Self-Play with a General Reinforcement Learning Algorithm". In: (2017). arXiv: 1712.01815 [cs.AI].

[SLS24] Xiaohou Shi, Jiahao Liu, and Yaqi Song. "BERT and LLM-Based Multivariate Hate Speech Detection on Twitter: Comparative Analysis and Superior Performance". In: *Artificial Intelligence and Machine Learning*. Ed. by Hai Jin, Yi Pan, and Jianfeng Lu. Singapore: Springer Nature Singapore, 2024, pp. 85–97.

[Sma+21] Christopher Small, Michael Bjorkegren, Timo Erkkilä, Lynette Shaw, and Colin Megill. "Polis: Scaling deliberation by mapping high dimensional opinion spaces". In: *Recerca: revista de pensament i anàlisi* 26.2 (2021).

[Sma+23] Christopher T. Small, Ivan Vendrov, Esin Durmus, et al. "Opportunities and Risks of LLMs for Scalable Deliberation with Polis". In: *ArXiv* abs/2306.11932 (2023).

[Ste+05] Jürg Steiner, André Bächtiger, Markus Spörndli, and Marco R. Steenbergen. *Deliberative Politics in Action. Analysing Parliamentary Discourse*. Cambridge: Cambridge University Press, 2005.

[Tan+24] Zhen Tan, Dawei Li, Alimohammad Beigi, et al. "Large Language Models for Data Annotation: A Survey". In: *ArXiv* abs/2402.13446 (2024).

[Tau+24] Amir Taubenfeld, Yaniv Dover, Roi Reichart, and Ariel Goldstein. "Systematic Biases in LLM Simulations of Debates". In: *ArXiv* abs/2402.04049 (2024).

[Tsa+24] Lily L. Tsai, Alex Pentland, Alia Braley, et al. "Generative AI for Pro-Democracy Platforms". In: *An MIT Exploration of Generative AI* (Mar. 2024). https://mit-genai.pubpub.org/pub/mn45hexw.

[Tsi+24] Dimitrios Tsirmpas, Ioannis Gkionis, Georgios Th. Papadopoulos, and Ioannis Mademlis. "Neural natural language processing for long texts: A survey on classification and summarization". In: *Engineering Applications of Artificial Intelligence* 133 (2024), p. 108231.

[Ulm+24] Dennis Ulmer, Elman Mansimov, Kaixiang Lin, et al. “Bootstrapping LLM-based Task-Oriented Dialogue Agents via Self-Talk”. In: *ArXiv* abs/2401.05033 (2024).

[Vec+21] Eva Maria Vecchi, Neele Falk, Iman Jundi, and Gabriella Lapesa. “Towards Argument Mining for Social Good: A Survey”. In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Ed. by Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli. Online: Association for Computational Linguistics, Aug. 2021, pp. 1338–1352.

[Vez+23] Alexander Sasha Vezhnevets, John P. Agapiou, Avia Aharon, et al. “Generative agent-based modeling with actions grounded in physical, social, or digital space using Concordia”. In: *ArXiv* abs/2312.03664 (2023).

[Wal+12] Marilyn Walker, Jean Fox Tree, Pranav Anand, Rob Abbott, and Joseph King. “A Corpus for Research on Deliberation and Debate”. In: *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC’12)*. Ed. by Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, et al. Istanbul, Turkey: European Language Resources Association (ELRA), May 2012, pp. 812–817.

[Wan08] Qiyun Wang. “Student-facilitators’ roles in moderating online discussions”. In: *Br. J. Educ. Technol.* 39 (2008), pp. 859–874.

[WC22] Yau-Shian Wang and Ying Tai Chang. “Toxicity Detection with Generative Prompt-based Inference”. In: *ArXiv* abs/2205.12390 (2022).

[WS07] Scott Wright and John Street. “Democracy, deliberation and design: the case of online discussion forums”. In: *New Media & Society* 9.5 (2007), pp. 849–869. eprint: https://doi.org/10.1177/1461444807081230.

[Xia+20] Yan Xia, Haiyi Zhu, Tun Lu, Peng Zhang, and Ning Gu. “Exploring Antecedents and Consequences of Toxicity in Online Discussions: A Case Study on Reddit”. In: *Proc. ACM Hum.-Comput. Interact.* 4.CSCW2 (Oct. 2020).

[XL24] Ruoyu Xu and Gaoxiang Li. *A Comparative Study of Offline Models and Online LLMs in Fake News Detection*. 2024. arXiv: 2409.03067 [cs.SI].

[Zha+16] Justine Zhang, Ravi Kumar, Sujith Ravi, and Cristian Danescu-Niculescu-Mizil. “Conversational Flow in Oxford-style Debates”. In: Apr. 2016, pp. 136–141.

[Zha+18] Justine Zhang, Jonathan P. Chang, Cristian Danescu-Niculescu-Mizil, et al. “Conversations Gone Awry: Detecting Early Signs of Conversational Failure”. In: *CoRR* abs/1805.05345 (2018). arXiv: 1805.05345.

[Zhe+24] Rui Zheng, Hongyi Guo, Zhihan Liu, et al. “Toward Optimal LLM Alignments Using Two-Player Games”. In: *ArXiv* abs/2406.10977 (2024).

[Zho+24] Hao Zhou, Chengming Hu, Ye Yuan, et al. “Large Language Model (LLM) for Telecommunications: A Comprehensive Survey on Principles, Key Techniques, and Opportunities”. In: *ArXiv* abs/2405.10825 (2024).

[ZN19] Qunyan Maggie Zhong and Howard Norton. “Exploring the Roles and Facilitation Strategies of Online Peer Moderators”. In: 2019.

Websites

[Ben+16] Susan Benesch, Derek Ruths, Kelly P Dillon, Haji Mohammad Saleem, and Lucas Wright. *Counterspeech on twitter: A field study. Dangerous Speech Project*. 2016.

[Con23] LangChain Contributors. *LangChain: Building applications with LLMs through composability*. GitHub repository. 2023. URL: https://github.com/langchain-ai/langchain (visited on Sept. 10, 2024).

[For14] World Economic Forum. *Moderation and Facilitation*. Jan. 2014. URL: https://www3.weforum.org/docs/WEF_Moderation_Facilitation_Briefing_2014.pdf (visited on Oct. 24, 2024).

[Gra08] Paul Graham. *How to Disagree*. Accessed: 2024-06-24. Mar. 2008. URL: https://paulgraham.com/disagree.html.

[Har24] Harvard Graduate School of Education. *Responding to Students*. Accessed: 2024-09-16. 2024. URL: https://instructionalmoves.gse.harvard.edu/responding-students.

[Ini17] Cornell eRulemaking Initiative. *CeRI (Cornell e-Rulemaking) Moderator Protocol*. Cornell e-Rulemaking Initiative Publications, 21. 2017. URL: https://scholarship.law.cornell.edu/ceri/21.

[Vin19] James Vincent. *Former Go champion beaten by DeepMind retires after declaring AI invincible*. Accessed: 2024-11-22. Nov. 2019. URL: https://www.theverge.com/2019/11/27/20985021/go-champion-retires-ai-invincible-deepmind-alphago-ai-artificial-intelligence (visited on Nov. 22, 2024).

List of Acronyms

API Application Programming Interface

OOP Object Oriented Programming

JSON JavaScript Object Notation

AI Artificial Intelligence

NLP Natural Language Processing

LLM Large Language Model

DL Deep Learning

ML Machine Learning

RL Reinforcement Learning

nDFU normalized Distance From Unimodality

SDB Socio-Demographic Background

SDL Synthetic Discussion Library

SDF Synthetic Discussion Framework

W.E.I.R.D. Western, Educated, Industrialized, Rich, and Democratic

IR Information Retrieval

List of Figures

List of Tables

List of Algorithms

Declaration

I hereby declare that this thesis titled "Mitigating Polarisation in Online Discussions Through Adaptive Moderation Techniques" submitted to the Natural Language Processing Groupof Athens University of Economics and Businessin partial fulfillment of the requirements for the degree of Master of Science in Data Science, is my original work, and it has not been submitted previously for any degree, diploma, or other qualification at any other university or institution.

This thesis combines the empirical fields of software engineering, data science, and natural language processing, and their interaction is evident throughout the research. All three of these disciplines have been of both scientific and general interest to me, and I hope that this work may serve as a foundation for future systems and experimentation procedures, contributing to further exploration in this area of research.

I affirm that all sources of information used in this thesis have been acknowledged, and I have not committed any form of plagiarism. Any assistance or contributions by others to the research and writing of this thesis, including any substantial editorial work, have been clearly indicated in the acknowledgments.

This work has been carried out under the guidance of Assistant ProfessorJohn Pavlopoulos.

Athens, Greece , October 2024

Dimitris Tsirmpas