Athens University of Economics and Business

ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ
ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS

Department of Informatics
Athens, Greece

Master Thesis
in
Data Science

Mitigating Polarisation in Online Discussions Through Adaptive Moderation Techniques

Dimitris Tsirmpas

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

Committee: Prof. Ion Androutsopoulos
Department of Informatics
Athens University of Economics and Business

Prof. Theodoros Evgeniou
Decision Sciences and Technology Management
INSEAD

October 2024

Dimitris Tsirmpas

Mitigating Polarisation in Online Discussions Through Adaptive Moderation Techniques

October 2024

Supervisor: Assistant Prof. John Pavlopoulos
Reviewers: John Pavlopoulos , Ion Androutsopoulos and Theodoros Evgeniou

Athens University of Economics and Business

Department of Informatics

Athens, Greece Patision 76
10434 and Athens, Greece

Abstract

Online discussion facilitation is crucial for discussions to flourish and prevent polarization and toxicity, which seem omnipresent. However, since it relies on humans, it 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 to substitute 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, although we prove that using socio-demographic backgrounds in LLM annotators does not meaningfully affect their judgments. Finally, we release a synthetic-discussion creation and annotation framework, three synthetic datasets resulting from our experiments as well as subsequent analysis and findings from these datasets[1].

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

Περίληψη

Οι διαδικτυακοί χώροι συζήτησης έχουν καταστεί ζωτικής σημασίας για τον υγιή διάλογο μεταξύ δισεκατομμυρίων ανθρώπων και για πολλές δημοκρατικές διαδικασίες. Ωστόσο, μαστίζονται από την τοξικότητα και την πόλωση. Οι σύγχρονες τεχνικές συντονισμού/διαμεσολάβησης (moderation/facilitation) διαλόγου είναι αποτελεσματικές στη βελτίωση της ποιότητας των συζητήσεων, αλλά απαιτούν ανθρώπινη συμμετοχή και, ως εκ τούτου, είναι δαπανηρές και μη επεκτάσιμες. Τα Μεγάλα Γλωσσικά Μοντέλα (ΜΓΜ, ή LLMs στα αγγλικά) μπορούν να παρακάμψουν αυτά τα προβλήματα αντικαθιστώντας εν μέρη τους ανθρώπινους διαμεσολαβητές, αλλά η ανάπτυξη συνθετικών διαμεσολαβητών είναι αργή, επιρρεπής σε σφάλματα και συνήθως απαιτεί ανθρώπινη συμμετοχή σε πειράματα, αυξάνοντας σημαντικά το κόστος της.

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. Επεκτείνουμε το σύστημα μας με τη δυνατότητα υποστήριξης επισημείωτων ΜΓΜ, για την αντιμετώπιση του προβλήματος της αξιολόγησης διαλόγων. Οι ψευτο-επισημειωτές αυτοί έχουν δικά τους κοινωνικο-δημογραφικά υπόβαθρα, έτσι ώστε να μοντελοποιήσουμε τη διαφωνία των ανθρώπινων αντιστοίχων τους, με βάση τις προκαταλήψεις τους. Τέλος, αναλύουμε την επίδραση διαφόρων παραγόντων στην τοξικότητα των συνθετική συζητήσεων, ως υποκατάστατη μετρική της ποιότητάς τους.

Δίνουμε δημόσια τον πηγαίο κώδικα του συστήματος, τρία συνθετικά σύνολα δεδομένων που αφορούν τις ίδιες τις συνθετικές συζητήσεις, τις επισημειώσεις τους και τα αμφιλεγόμενα σχόλια σύμφωνα με τους διάφορους επισημειωτές ΜΓΜ. Η διατριβή εμπεριέχει επίσης και πειράματα, γραφήματα και στατιστικούς ελέγχους που αποδεικνύουν τα συμπεράσματά μας. Συμπεραίνουμε ότι *οι συνθετικές συζητήσεις που διεξάγονται αποκλειστικά μέσω χρηστών ΜΓΜ, μπορούν να βοηθήσουν στον εντοπισμό μοτίβων συμπεριφοράς ανάλογα με την επιλεγμένη τεχνική διαμεσολάβησης*. Από την άλλη, *διαψεύδουμε* την επίδραση των κοινωνικο-δημογραφικών υποβάθρων στην επισημείωση δεδομένων.

Acknowledgements

I would foremost like to thank the professors John Pavlopoulos , Ion Androutsopoulos and Theodoros Evgeniou for their help, insights and encouragement in every stage of this thesis. Special thanks to Assistant Prof.John Pavlopoulos for personally supervising the thesis and encouraging me to bring it to fruition until the very end. I would also like to thank the colleagues and researchers at the Athena Research Center, whose input helped steer the project towards more productive directions. Lastly, I am grateful to my family, whose constant support has greately aided me in all my endeavors.

Contents

Introduction

1

1.1 Motivation and Problem Statement

Large Language Models (LLMs) have revolutionized the field of Natural Language Processing (NLP) since their introduction 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, have led to many established companies, startups and research groups around the globe to scramble and identify useful use-cases for this novel technology [Had+; 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 facilitators"*, people whose responsibility is to uphold the rules of the discussion and discipline users ("discourse moderation"). 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]. However, 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, configurations (e.g. which model family, whether to use pretrained or finetuned models, ...) by the use of robust experiments with human subjects can be similarly expensive on the researchers' side.

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 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. Both source code and datasets can be found at the project's repository [1].

Our framework further incorporates automated LLM-based annotations of these synthetic discussions, allowing for an inexpensive comparison of the effects of various factors. By using annotators with different SDBs, we assess whether different LLM personalities influence the annotation process. We experiment with various prompt strategies and configurations to evaluate how they affect conversation quality, using toxicity as a proxy. Finally, we analyze the content of the discussions alongside the LLM annotations and generate three synthetic datasets that include the discussions, their annotations, and the inter-annotator agreement.

Alongside the creation of this framework to aid in experimentation for online LLM facilitation, we try to answer the following two questions: **Q1:** Can LLMs convincingly argue against each other when supplied with only a controversial topic and differing SDBs? **Q2**: Do LLM annotators change their behavior according to different 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) delves into 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 hypothesized and applied facilitation tasks, 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)

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

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 and strategies used in both the synthetic creation and annotation loops. 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 conversations.

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 the experiments. Finally, we discuss the possible research avenues this thesis opens for future exploration.

2 Background and Related Work

2.1 Background

2.1.1 How and why do humans argue?

Collective deliberation and decision-making has been long hypothesized, and proven, to yield better results than those performed by individuals [MG98; Sch+06]. This idea has often been expressed by the phrase "the group is better than the sum of its parts".

Social science research often attempts to categorize distinct tactics in arguments. Graham [Gra08] propose a hierarchy of disagreements, ranging from name-calling, to refuting the central point of an argument. While a convenient framework, it has not been verified empirically [DSV22]. 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 [DSV22]. There are many attempts at refining the original hierarchy, such as the one presented by Benesch et al. [Ben+16].

Disagreements and toxicity are a natural part of human dialogue, which however often lead to the discussion failing. De Kock et al. [DSV22] 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 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 conversations differ greatly from offline (face-to-face) conversations. 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 conversations 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, facilitation of effective communication, 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].

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, while the research of Zhang et al. [Zha+16] claim to investigate how ideas flow between groups holding and discussing different views, and while their insights are doubtlessly important, the authors end up investigating what wins an argument, and 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.

In a system aiming to facilitate discussion and find common ground among participants, such thinking will inevitably lead to a platform designed instead to provoke arguments, attempts at attacking and antagonizing the other side and to foster a culture of "winning" discussions by any means necessary. The phenomenon is also mentioned in Karadzhov et al. [KSV21], alongside the fact that most existing datasets involve only two participants, whereas deliberating platforms usually involve group thinking and deliberation.

2.1.4 Large Language Models

LLMs are sophisticated Artificial Intelligence (AI) computational models capable of textual language generation by training on vast amounts of text data largely scrapped from the wider Internet. LLMs are based on the Transformer architecture [Vas+17], after it was widely adopted in numerous models undertaking many NLP tasks. Without going into the history of how these models came to be, it is sufficient to say that LLMs used next-word-predictions to fulfill general tasks given by user-defined prompts. Because of their extensive size, complexity and pretraining, these models managed to compete with previous specialized models in multiple tasks such as Topic Classification, Sentiment Analysis, Text Summarization [Tsi+24], as well as specialized annotation tasks [Tan+24]. Even more than that, they also proved capable of executing general tasks, leading to their worldwide use as personal assistants, automated systems, chatbots, and many more such roles.

Another interesting property of LLMs 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, if not impossible; indeed, should this have been not been the case, human-involved studies would have become redundant.

Lastly, 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.

2.2 Related Work

2.2.1 LLM self-training

Using unsupervised finetuning by means of initiating conversations where the participants are controlled by the same LLM instance, has become an area of intense research into LLMs. 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. However, not all self-talking approaches use RL or an adversarial scenario, nor are they used exclusively in the context of jailbreak prevention.

Abdelnabi et al. [Abd+23] 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's success is quantified by the so-called scores of the parties and agreement thresholds, which need to be surpassed in order for an actor to be able to select an 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 framework itself is very difficult for most LLMs; preliminary results show that most of the time, GPT-3.5 and smaller models fail, while GPT-4 and other state-of-the-art models underperform.

Another interesting idea is "Self-play" a RL technique 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, the highest-profile being chess by the Alpha-Zero model in 2017 [Sil+17]. 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. More 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 used a 30 billion parameter MosaicAI [Tea23] model for the client, and a 7 billion parameter model of the same family for the agent (since the agent is inherently greatly restricted). Only the agent model is finetuned. The researchers demonstrate that self-talk can indeed be used to improve LLMs, given enough finetuning and rigorous filtering of input data. What is important to this thesis, however, is that it provides a practical demonstration that LLMs conversing with each other can produce quality conversations when applied in a structured setting, even if they are ultimately not used for model finetuning.

LLM self-play is hypothesized to work in discourse facilitation tasks. Small et al. [Sma+23]s 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.*

It is important to note that conversations 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. Their system is also interactive, allowing the end-user to experiment by changing community rules or individual personas on a local level and observing the changes in the conversations (for example, what would be the impact on the conversation if this comment was made by a troll?). Thus, social simulacras can act as a form of prototyping for internet communities. The researchers show that appropriately prompted LLMs using generated personas are adequate at mimicking human users, their posts being generally indistinguishable by the mirrored actual communities to human annotators.

Finally, Lambert et al. [Lam+24] follow the work of Bai et al. [Bai+22] and create a self-regulating conversation generation framework. Specifically, they use a set of given topics by Castricato et al. [Cas+24], various principles fundamentally based on human rights, and define various conversation goals (help a user create an email, an essay, perform language translation etc.). An LLM then generates a plan (system prompt) for the conversation and begins generation while checking if at any point the principles have been violated. In that case, it generates a critique on why the conversation failed. The models are encouraged to violate the goals of the conversation for the sake of data quality. It is worth noting that, the study failed to find any trends between principles, goals and the generated text.

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 including sociodemographic information in LLM prompts can in some situations greatly increase their performance in various NLP tasks. Specifically, they show that changing sociodemographic information significantly influences classification results (which is also observed between humans of different social and demographic groups), although the results are contingent on the prompt structure, model family and model size, in non-obvious ways. Large models (containing more than 11 billion parameters) can often leverage this information, primarily using combinations, instead of individual traits, although they can not use them as explanatory variables.

However, sociodemographic prompting does include caveats. Asides from the non-existence of robust prompting templates and models that can reliably leverage sociodemographic information [Bec+24], skepticism exists concerning stereotypical biases [CDJ23; Des+23] as well as models having a large bias towards responses from Western countries, and the unavailability of relevant datasets concerning languages other than English [San+23a; Dur+24; San+23b] as cited by Beck et al. [Bec+24]. Furthermore, Aher et al. [AAK23] cite SD "distortions" as a recurring problem, where the model's responses and behavior deviate significantly from what is expected of a human bearing the same SDB information. The researchers point to an example where a LLM pretending to be an average human could include in its response something as specific as the melting point of aluminum.

2.2.3 LLMs as discourse facilitators

LLMs can 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 the traditional system, a facilitator would present the participants with a summary of a key standpoint or worldview they presented as he 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 the active discussion to test hypotheses about the different groups' opinions, which is especially useful in finding common ground. It is hypothesized that using this procedure with an LLM may yield faster convergence to common ground and model understanding of the opinions of the participants. Another interesting area of interest is using LLMs to directly produce opinions [Sma+23] at the start of the dialogue (called "seed opinions" in the original papers), which the authors claim have a significant impact on the course of the discussion. However, Karadzhov et al. [KSV21] demonstrate that synthetic data (based on their own pretrained LLM) are less convincing than retrieval-based, or even random selection of phrases from similar discussions, both on many metrics, and by human opinion. This phenomenon is more prevalent on issues which necessitate advanced vocabulary and reasoning.

Al-Khatib et al. [Al-+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. Thus, a LLM moderator 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 the researchers define the goal of a deliberative discussion differently than the one used by this paper and defined 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 by 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 (called "problematizer" in the original paper) 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, LLMs are well positioned to tackle traditional NLP problems relating to online discussions; namely machine translation (in order to allow marginalized and minority groups to contribute) [Tsa+24], hate-speech [Nir+24; SLS24], toxicity [KQ24; WC22] and fake-news [Liu+24; XL24; 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 and has a direct correlation with metrics used in NLP tasks [Wac+17]. x

De Kock et al. [DSV22] 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 respective 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 LLMs performance on actual conversations, we still lack a way of quantifying the quality of the *synthetic* dialogues. Ulmer et al. [Ulm+24] propose a series of automated evaluation metrics for synthetic dialogues. "Dialogue Diversity" 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. "Subgoal completion" calculates the ROUGE-L score between the LLM's response to a question and predefined utterances in the LIGHT [Urb+19] dataset, containing fantasy quests, to determine decisions taken by the LLM; these are then compared to a graph mapping of all possible paths in the dataset, and are given a completion score according to how close the LLM was to an ending. Finally, "Character Consistency" measures how much the LLM stays in-character and is evaluated by a finetuned DeBERTa [HGC23] model.

2.2.5 Risks and Challenges

First of all, we feel compelled to echo the author's warnings in [Sma+23]. Synthetic data and conversations 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] outline several known weak points in LLM usage for moderation; LLMs suffer from bias, hallucinations, are vulnerable to prompt injection attacks, and have their own political leanings (with most trending towards progressive ideas). 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 doesn't discriminate against users who don't belong in 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; personal opinions, facts and fake news are often bundled together.

Lastly, training generative models, and more specifically LLMs, on their own data most often leads to the model collapsing [Ale+23; Shu+24] as cited by Ulmer et al. [Ulm+24]. Additionally, even when not trained on their own data, LLMs tasked with creating dialogues often generate low quality, off-topic and generally useless data [Ulm+24]. Their experiments show that at many points the conversation collapses with the models going off-script, rambling or ending 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 generated topic diversity [Lam+24].

2.2.6 Related Datasets

One of the most frequently used datasets for goal-oriented discussions is the Wikipedia Disputes dataset [DV21], which contains discussion from the 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. [DSV22] provide the "WikiTactics" dataset, a dataset built on the former, which provides annotations based on the tactics employed in each utterance in the context of each dispute. Hua et al. [Hua+18] expand on the Wikipedia Disputes dataset, creating "WikiConv", encompassing all contributor conversations on Wikipedia. The dataset is novel in that it includes metadata concerning edits, deletions and other actions on the comments themselves, allowing for further accurate analysis of these conversations. This approach was followed by Al-Khatib et al. [Al-+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 2400 turns labeled with discourse acts, 7437 turns labeled with relational connections between utterances, 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 conversation derailment datasets are also available, albeit in relatively small numbers. Zhang et al. [Zha+18] provide a curated dataset of 1270 conversations with an average length of 4.6 comments each, featuring derailed conversations. Chang and Danescu-Niculescu-Mizil [CD19] provide two datasets relating to discussion derailment, the first expanding on the previous dataset with a total size of 4188 conversations and a larger discussion length, while the second is sourced from the "Change My View" (CMV) Subreddit, featuring $600,000$ conversations, 6842 of which necessitated moderator intervention.

One of the few datasets containing group discussions is the "Deli Data-Deliberation Dataset" [KSV21], which includes 500 group discussions, and is annotated by both metadata and an objective measure of correctness based on whether a given task (e.g. solve a given puzzle) is solved. The metadata are comprised of three categorizations which concern whether a statement exists to provoke discussion or share information, which specific role it plays within the context of the discussion, and additional information on specific phenomena. Of course, this dataset quantifies quality as success in a specific task which, while proven to work in other out-of-domain tasks, may not generalize well to platforms where there is no defined task.

Synthetic-only dialogue datasets are exceedingly rate in literature. [Lam+24] provide a dataset containing $108,000$ sentences generated by different models, using a topic, subtopic and goal for each conversation. They also publish a sister dataset containing the LLM annotations for why the conversation violated the stated policies of the discussion.

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 of dialogues through LLMs. In this section we explain in detail the initial requirements for this framework and why commercially-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

The requirements for the SDF were not obtained by standard requirement solicitation procedures. Thus, no formal document detailing them exists. Instead, they were iteratively solicited during weekly meetings with the wider 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 conversation 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 conversation.
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.

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.
4. The system must support large-scale data annotation.
5. The system must support a diverse and flexible array of annotation criteria.

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, as they are industrial-grade frameworks, meant for a diverse set of business use-cases, making their API convoluted. Of course, this could be circumvented by employing the Adapter pattern [Gam+95]. The problem then would be that their internal components frequently necessitate computer resources (dedicated RAM, GPU VRAM e.t.c.) which, for a smaller application such as ours, will most likely not be used to their fullest.

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+23]. Figure 3.1 shows a simplified version of a conversation involving only two users

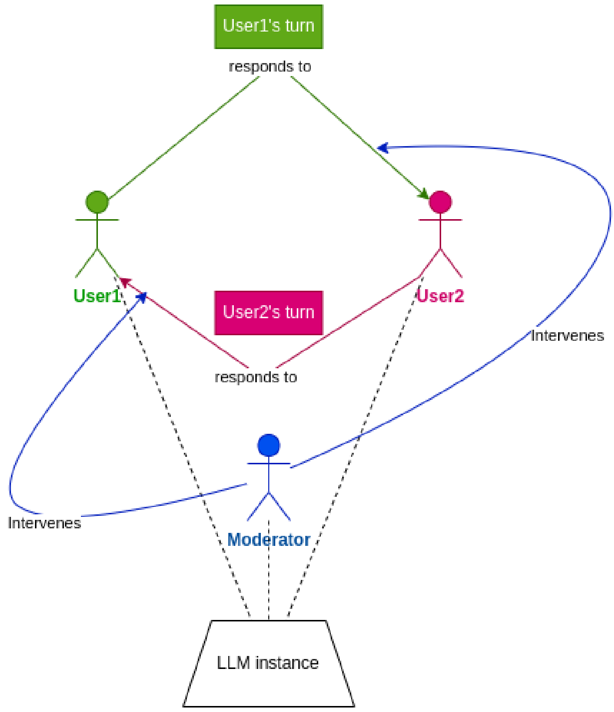

Fig. 3.1: The conversation loop on which the SDF operates. Can be generalized for N users and 0 or 1 moderators.

and the moderator. A more general overview of the discussion creation loop can be found in Algorithm 1.

The two Users and the Moderator are all controlled by the same LLM instance; we only change the system prompt when each takes its turn to speak. The prompts are comprised of five parts:

- **Name**, the name of the actor, used for other actors to refer to him in-conversation.
- **Role**, the role of the actor within the conversation (user, moderator).
- **Attributes**, a list of actor attributes, primarily used for giving the actor an SDB.
- **Context**, information known to all users.
- **Instructions**, potentially unique to each actor.

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 conversations. This could be achieved by using

Algorithm 1 Synthetic Dialogue Creation algorithm

```
Input users, maxTurns, historyLength
Output the conversation 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])

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

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 Machine Learning (ML)-based instead of Deep Learning (DL)-based [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. LLM inference is practically constant-time with a fixed input length, since adding a new annotation metric would only impose a computational penalty equal to the output's increased number of tokens - which is negligible.

Using LLMs as annotators imposes both a challenge and an opportunity, since annotations are no longer objective (unlike traditional ML and even DL models, we can't explain a LLM's decision). 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.
- Lean into the subjectiveness of LLM decision-making and use many LLM annotators, each with a different SDB, then compute inter-annotator agreement.

In this thesis, we use the second option.

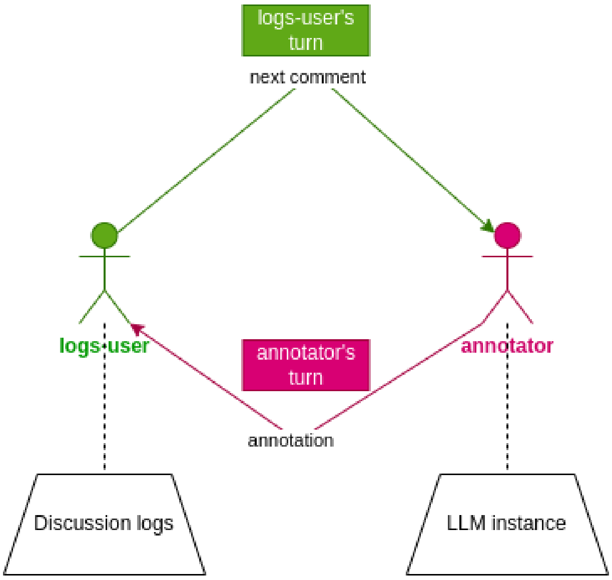

Fig. 3.2: The annotation loop on which the SDF operates. Note the purposeful similarity of the function to Figure 3.1.

We re-use the conversation paradigm of Section 3.2.1 to facilitate annotation. One pseudo-actor is the system, which outputs comments made in a conversation 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 of 4 for the annotator; in other words, for each comment, the annotator can see the 4 preceding comments of the conversation. The annotation loop is succinctly demonstrated in Figure 3.2 and in Algorithm 2.

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
```

3.3 Prompt Design

3.3.1 Defining Policy & Environment

Our user instruction prompts in general were designed with the following criteria in mind:

- Many people are unwilling to change opinions during online discussions.
- Personal attacks are common [DSV22].
- 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.

Thus, user prompts are designed to promote toxicity, controversy and confrontation.

On the other hand, our moderator prompts were largely based on the Cornell eRulemaking moderator manual [Ini17] which, among others, emphasizes the following guidelines:

- The moderator must remain neutral and impartial.
- Responses should be briefly reflected upon before being posted.
- 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.

3.3.2 "Moderation Game" prompts

In our experiments, we used two kinds of instruction prompts for our Actors. One is a "standard" prompt summarizing the guidelines above. 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+23]. 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.

Thus, the user scores were defined as:

- 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

and their moderator equivalents as:

- Intervene: -1 points
- Threatened ban: -1 points
- Intervention led to better behavior: +3 points
- Banned a participant: -5 points

3.3.3 Annotator prompts

The annotator prompt consists of the following parts:

- The SDB prompt.
- An instruction prompt, in this case geared towards toxicity classification. This part however can be replaced to make the model output any combination of annotations.
- A list of examples with varying toxicity (few-shot learning).
- The output prompt.

Due to limitations in context window length, the prompt only contained basic information and only a few examples.

3.4 Implementation

3.4.1 Synthetic Discussion Library

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, contains 4 distinct modules, and is based on Object Oriented Programming (OOP) principles.

Each of these modules contains classes and supporting code for a specific function. In brief:

- **models.py** holds Adapter classes [Gam+95] which enable the framework to uniformly access almost any LLM instance regardless of type (as long as a suitable subclass is created).
- **actors.py** which holds Wrapper classes [Gam+95], containing Model Adapter classes from **models.py** and provide them with prompt templates.
- **conversation.py** uses Actor classes in order to execute and serialize the conversation.

The library additionally provides the **annotator.py** and **util.py** modules, which are self-explanatory.

3.4.2 Framework entry-points

The framework provides a variety of APIs to access the SDL from the more 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, 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.

3.4.3 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 analyses.

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 e.t.c.)

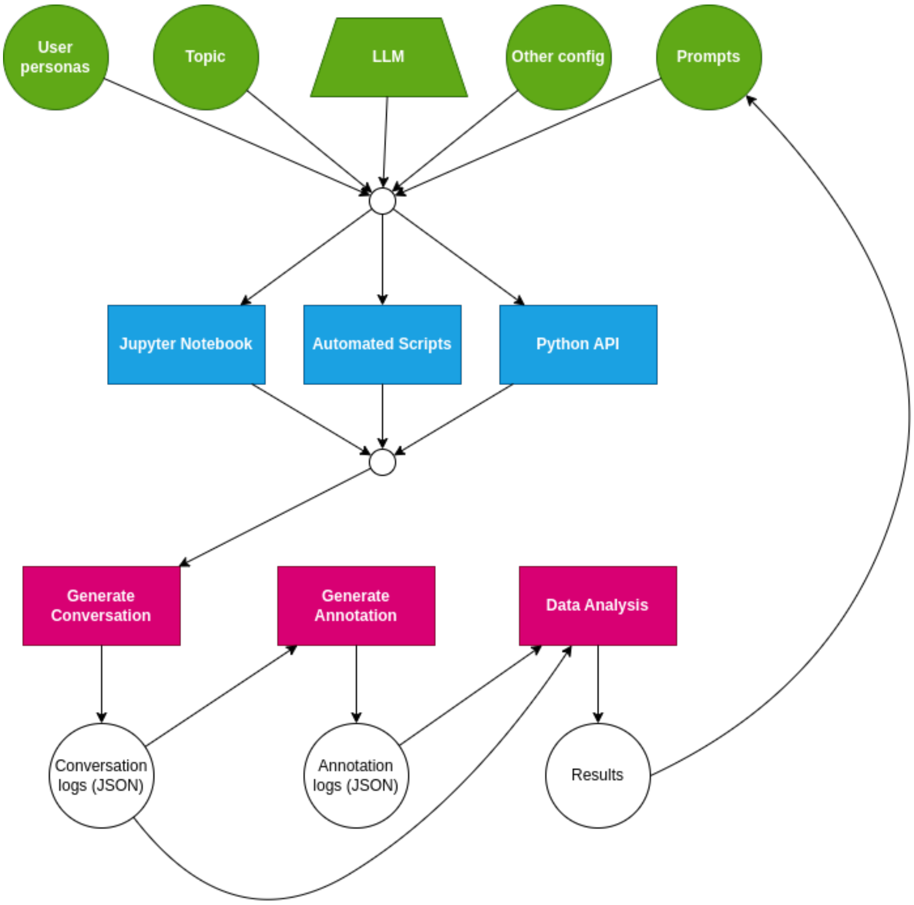

Fig. 3.3: An abstract view of the SDF. Green shapes represent various configurations, blue shapes entry points (see Section 3.4.2), pink ones processes delegated to the SDL, and **white** ones exported data.

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.

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

3.4.4 Technical Details

The LLM used is the LLaMa2-13B GGUF quantized version. We use the `llama_cpp` python library to load and interact with the model. 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 conversations 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 4.1. 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 4.2.

3. Select whether a moderator is present.

4. Provide actors with the global context.

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

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

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, except for "freezing" conversation-specific configurations. As such, we use a window of 4 comments and a conversation length of 3 Round-Robbin 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.

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 some 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 thats a no brainer hillary clinton gave huma abdein a security clearance when she has ties to a known terrorist group the muslim brotherhood, and her mother runs an anti american news paper in the middle east, debbie washed up crook shultz got the awan family security clearances and they were recent immigrants, had absolutely no IT experience and possible ties to terrorist groups in pakistan. its pretty clear our liberal ran 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 wont get off their behinds to do the damn job they are hired to do and 90% of government employees allow their personal and political agenda's 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 with out union.
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.

Tab. 4.1: Controversial topics used as seeds for the simulated conversations. Excerpts selected from Pavlopoulos and Likas [PL24].

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

Tab. 4.2: SDBs given to LLM users during the production of synthetic dialogues.

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.

4.1.2 Automated Dialogue Annotation

For each produced synthetic dialogue, we pick one out of the annotator SDBs present in Table 4.3. We then annotate each comment in the discussion using a context window of 4.

For the purposes of analyzing inter-annotator agreement we use the normalized Distance From Unimodality (nDFU) [PL24], a measure used to evaluate the dispersion of opinion distributions in a data set. DFU, or Discrepancy From Unimodality, is originally defined as the maximum difference between consecutive bins in a histogram representing the distribution of opinions. It is based on the assumption that opinions can be divided into several categories, and each category's frequency is recorded in the histogram. To normalize DFU, it is divided by the frequency of the most common opinion category (the mode). This step scales the measure so that the final score falls between 0 and 1, where 0 represents a unimodal distribution (indicating consensus around a single opinion) and 1 indicates maximum discrepancy in a multimodal distribution (indicating significant variation in opinions). The normalization helps to make the measure more interpretable, especially when comparing different distributions. Without normalization, the DFU score could be difficult to interpret because it is unbounded and dependent on the scale of the original data.

4.2 Produced Datasets

We produce three synthetic datasets:

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.

- The **Synthetic Dialogues Dataset**, containing the logs of the conversations, as well as rich metadata such as the prompts used and the conversation-specific configurations.
- The **Automated Annotation Dataset**, containing the annotations for each comment in each synthetic conversation. 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 conversation 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`. 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

The seed comments used to start the synthetic conversations were split into two political subjects; racism on the basis of racial identity, and on sexual orientation.

Table 4.5 presents the expected and actual behavior of synthetic users from different SDBs across the two polarized topics. The observations reveal notable deviations from expected behaviors.

For W.E.I.R.D. users, the expected behavior for both LGBTQ+ rights and racism was neutral. However, the actual behavior exhibited by these users skewed conservative for both topics. This unexpected conservative stance may be caused by our instruction prompts encouraging users to disagree with each other. The truly unexpected behavior was that

LLM user SDB	Topic	Expected behavior	Actual behavior
W.E.I.R.D.	LGBTQ+ rights	Neutral	Conservative
	Racism	Neutral	Conservative
LGBTQ+	LGBTQ+ rights	Progressive	Progressive
	Racism	Progressive	Progressive
African American	LGBTQ+ rights	Neutral	Progressive
	Racism	Progressive	Progressive

Tab. 4.5: Expected and observed behavior of synthetic users during our experiments by SDB.

African American users, who were always the first to speak, always adopted a progressive stance in topics concerning the LGBTQ+. This is not a behavior which necessarily matches with reality, since traditionally, many African Americans may adopt conservative stances [Loc13; MR13]. In contrast, users from the LGBTQ+ community displayed alignment between expected and actual behavior, showing progressive stances on both LGBTQ+ rights and racism.

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 stereotypical stances (as was the case with the African American synthetic users).

4.3.2 Impact of prompting strategies and moderator presence

In this section, we investigate the following hypothesis: **Different prompting strategies and moderator presence influence the toxicity of the conversations with identical topic and configuration**. The strategies used are the ones described in Section 3.3.

Figure 4.1 shows the mean toxicity for each prompting strategy, with or without moderator, for each annotator SDB. The red line shows the expected observed toxicity of the conversation (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. Finally, we note that some annotator SDBs generally gravitate towards different annotation scores; progressive SDBs such as the "African American" (green) and "LGBTQ+" (gray) annotators are more likely to mark a comment as toxic, than more conservative ones such as the "Blue Collar Worker" (brown).

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 Dunns 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 influences the level of toxicity..
- The prompting strategy significantly influences the toxicity level. The "Moderation Game" prompt keeps the conversation much more civil than the vanilla prompting strategy.
- The presence of a moderator does not influence the toxicity of the conversations 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 conversations.
- **Hypothesis 2**: The LLM users under the "Moderation Game" prompt are cautious of moderator action regardless of their presence. This hypothesis is reinforced by the fact that the LLM users 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.*

4.3.3 Impact of SDBs in LLM annotators

In this section, we test the following hypothesis: **Different LLM annotator SDB prompts influence the toxicity annotations *for the same* given conversation, in significant qualitative and statistical terms**.

We foremost check whether disagreement exists between the various annotations. Figure 4.3 shows the nDFU[PL24] scores for each synthetically created comment. The majority

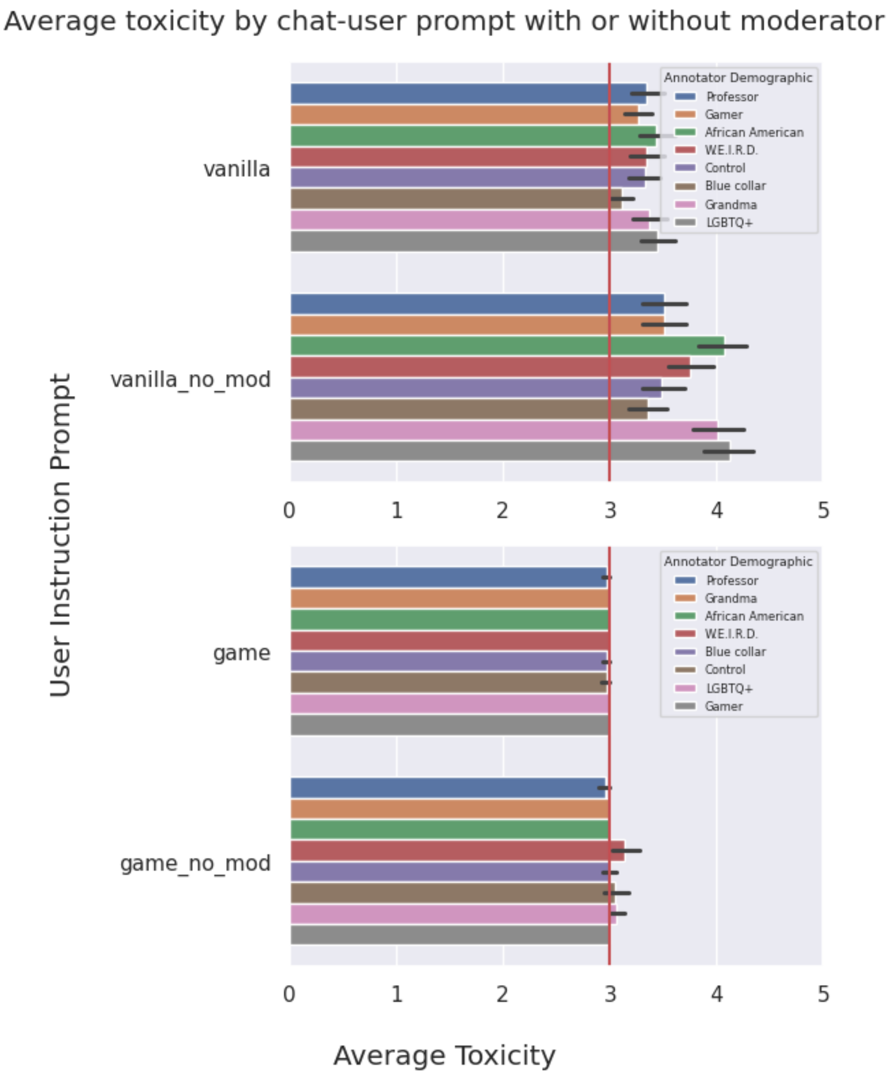

Fig. 4.1: Mean toxicity by prompting strategy and moderator presence, per annotator SDB.

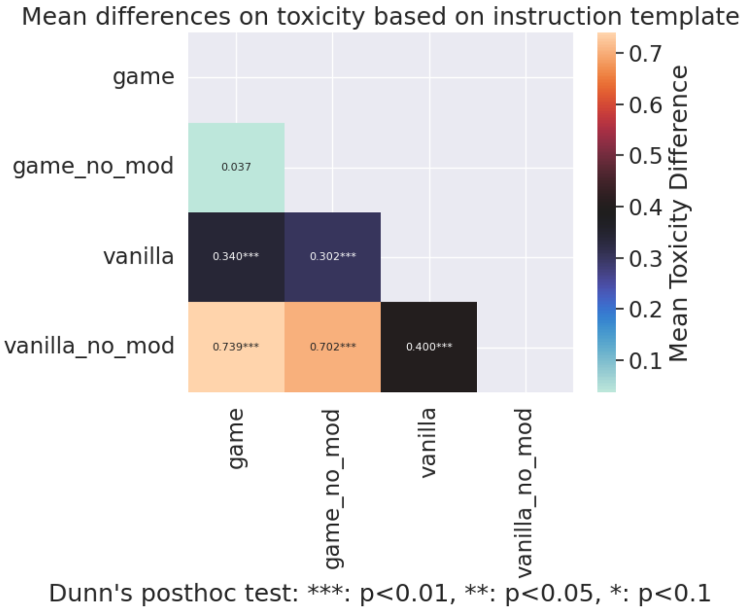

Fig. 4.2: Mean annotation difference between each strategy/moderator presence. Each comparison is accompanied by Dunn's posthoc test for multiple comparisons in the form of significance asterisks.

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 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 and, interestingly, specifically between some progressive-leaning (African American, LGBTQ+) and conservative-leaning (Blue collar) SDBs. *However, this pattern does not hold for all SDBs*, for instance between the "African American" and "Gamer" prompts where no significant deviations are observed. Finally, even though there exist statistically significant deviations,

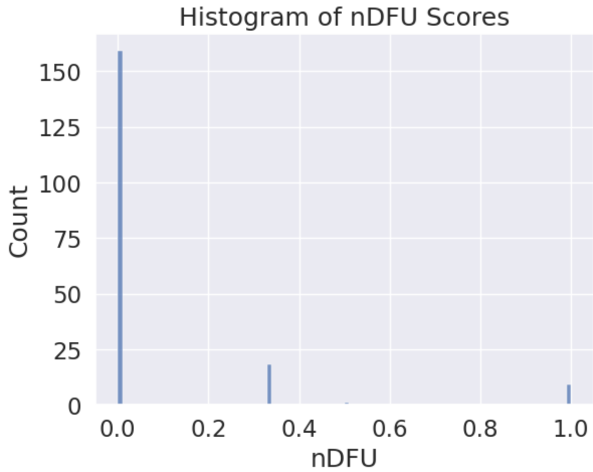

Fig. 4.3: nDFU [PL24] scores for each comment. More is larger disagreement between the annotators.

LLM user SDB	Expected cluster	Actual cluster
Blue Collar	Conservative	1
Grandma	Conservative	2
LGBTQ+	Progressive	2
W.E.I.R.D.	Neutral	3
African American	Progressive	3
Professor	Neutral	3
Control	Neutral	3
Gamer	Conservative	3

Tab. 4.6: Expected and observed clusters of synthetic annotators during our experiments by SDB.

these differences are not considerable. Indeed, the largest deviations only appear in the range of ± 0.3 mean toxicity annotation difference.

From Figure 4.5 we can infer the existence of behavioral clusters for each SDB. Table 4.6 showcases the expected and actual clusters for each SDB as inferred by the mean difference of annotations showcased in Figure 4.5. Note that our expected behavioral model is completely different from the actual annotations, indicating that SDBs have failed to model human annotators.

In order to further 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.

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)

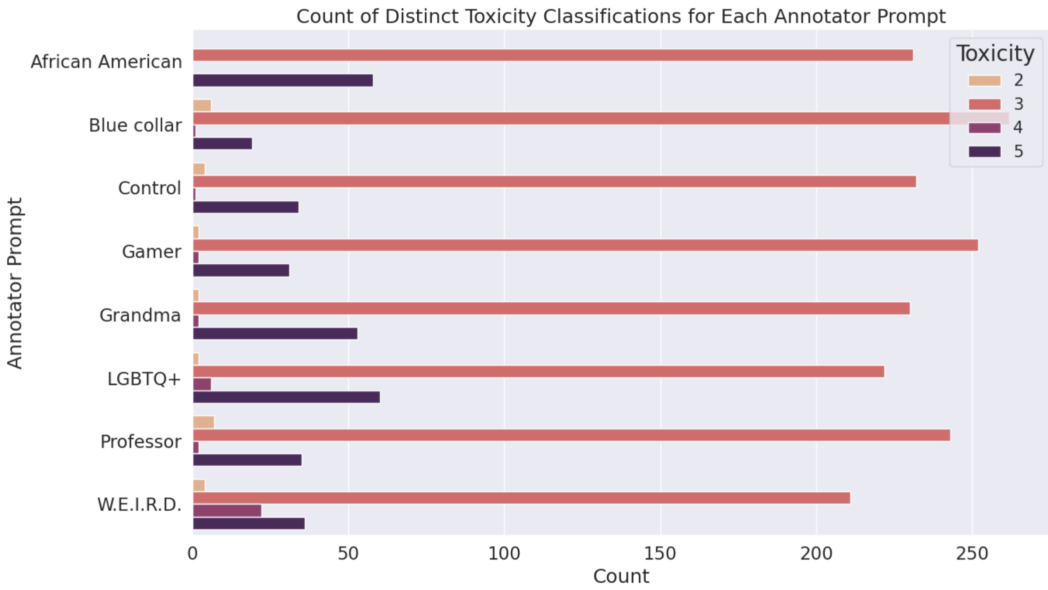

Fig. 4.4: Toxicity annotations by annotator SDB prompt. Note the high preference towards group 3 ("moderately toxic") and that significant deviations only occur between groups 4 ("very toxic") and 5 ("extremely toxic").

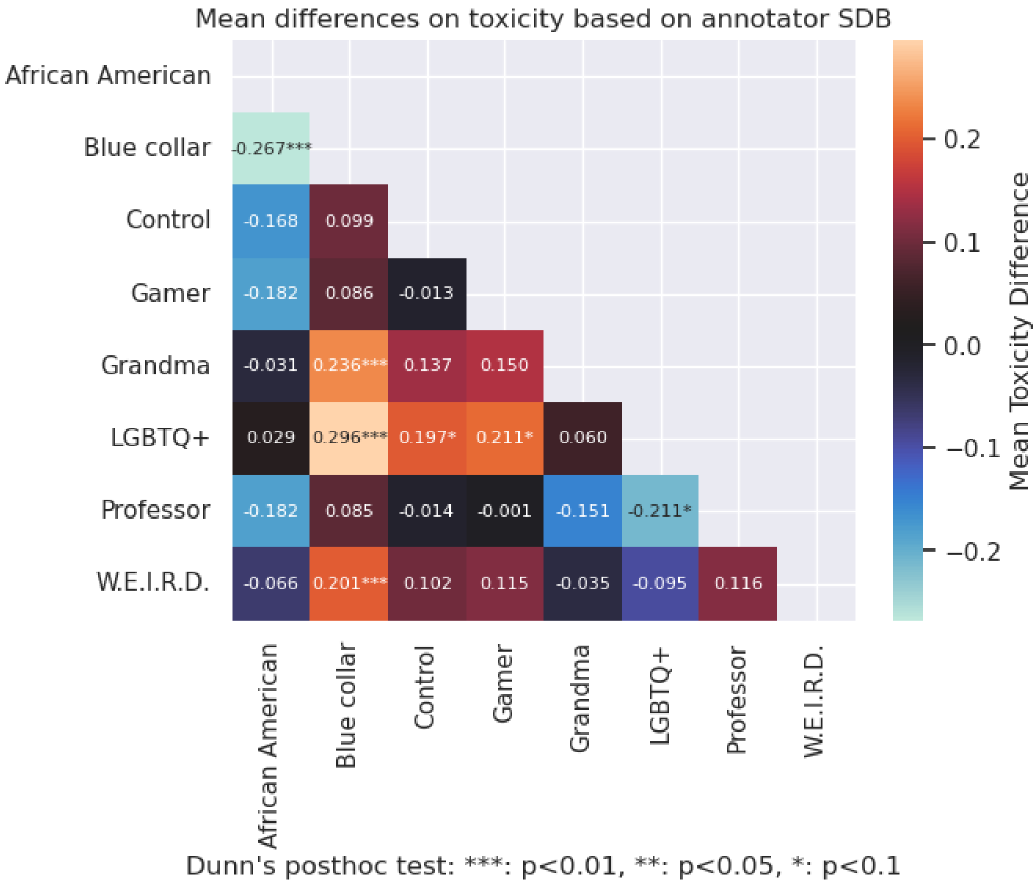

Fig. 4.5: Mean annotation difference between each annotator SDB. Each comparison is accompanied by Dunn's posthoc test for multiple comparisons in the form of significance asterisks.

within a feature. Then, feature G explains the polarization in X if $nDFU(X) > 0$, but $nDFU(X_i) = 0, \forall i \in G$. This criterion is intuitive and explainable, but does not cover cases where $nDFU(X_i)$ is close, 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". The algorithm presented in Algorithm 3 implements a statistical test for aposteriori unimodality to evaluate whether a given feature explains the observed polarization in a set of grouped annotations. First, the global nDFU is computed by concatenating all the annotations. Next, the nDFU is calculated individually for each factor of the selected feature. A Wilcoxon signed-rank test is then performed to assess whether the nDFUs of all factors are statistically indistinguishable from zero. Since annotation data typically does not follow a normal distribution and is often limited in quantity, the non-parametric Wilcoxon test is chosen for its robustness. 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.

Algorithm 3 Our proposed Aposteriori Unimodality Test

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

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

$ndfus \leftarrow \{\}$
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 selecting the "instruction prompt / moderator presence" feature, which curiously also does not explain the polarization between annotators ($p = 0.9375$). Thus, changing the instruction prompt and adding/removing the moderator meaningfully changes the toxicity level of the conversation, but the annotators stay consistent in the way toxicity changes.

Finally, we manually check the Controversial Discussions Dataset (see Section 4.2). We showcase two interesting comments in Table 4.7. The comment on the left is clearly conservative-leaning, while the one on the right, progressive-leaning. Both comments are

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

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, display that **annotators seem to differ in their priors**.

5 Conclusions & Future Work

In this thesis, we researched the practical 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; vanilla instruction prompting as well as framing the discussion as a competitive, scorable game. We then used this framework to generate three 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 the synthetic dataset, we found that the presence of a moderator 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 proved that using different SDBs in LLM annotators yields no significant qualitative difference, and that any difference can be attributed to a change in priors, as opposed to reacting differently according to the content and context of the synthetic messages.

Future work should expand on making synthetic conversations more realistic, ideally rendering them indistinguishable from human online conversations. 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 moderating techniques, interventions and model instances on conversation 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.

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+23] Sahar Abdelnabi, Amr Gomaa, Sarath Sivaprasad, Lea Schonherr, and Mario Fritz. "Cooperation, Competition, and Maliciousness: LLM-Stakeholders Interactive Negotiation". In: 2023.

[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.

[Al-+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.

[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), pp. 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).

[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-Niculescu-Mizil. "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.

[DSV22] 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.

[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]`.

[DV21] 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.

[Gam+95] Erich Gamma, Richard Helm, Ralph Johnson, and John Vlissides. *Design patterns: elements of reusable object-oriented software.* Pearson Deutschland GmbH, 1995.

[Had+] Muhammad Usman Hadi, al tashi, Rizwan Qureshi, et al. “A Survey on Large Language Models: Applications, Challenges, Limitations, and Practical Usage”. In.

[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: 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).

[Loc13] Brad Lockerbie. “Race and religion: Voting behavior and political attitudes”. In: *Social Science Quarterly* 94.4 (2013), pp. 1145–1158.

[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`.

[MR13] Brian D McKenzie and Stella M Rouse. “Shades of faith: Religious foundations of political attitudes among African Americans, Latinos, and Whites”. In: *American Journal of Political Science* 57.1 (2013), pp. 218–235.

[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 Park, Lindsay Popowski, Carrie Cai, et al. “Social Simulacra: Creating Populated Prototypes for Social Computing Systems”. In: Oct. 2022, pp. 1–18.

[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.

[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.

[Sch+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.

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

[Shu+24] Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, et al. “The Curse of Recursion: Training on Generated Data Makes Models Forget”. In: (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).

[Tea23] The MosaicML NLP Team. “Introducing MPT-7B: A New Standard for Open-Source, Commercially Usable LLMs”. In: *Mosaic AI Research* (May 2023).

[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).

[Urb+19] Jack Urbanek, Angela Fan, Siddharth Karamcheti, et al. "Learning to Speak and Act in a Fantasy Text Adventure Game". 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. 673–683.

[Vas+17] Ashish Vaswani, Noam M. Shazeer, Niki Parmar, et al. "Attention is All you Need". In: *Neural Information Processing Systems*. 2017.

[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).

[Wac+17] Henning Wachsmuth, Nona Naderi, Yufang Hou, et al. "Computational Argumentation Quality Assessment in Natural Language". In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*. Ed. by Mirella Lapata, Phil Blunsom, and Alexander Koller. Valencia, Spain: Association for Computational Linguistics, Apr. 2017, pp. 176–187.

[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”. In: 2024.

[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).

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).

[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.

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

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 Department of Informatics of Athens University of Economics and Business in 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 Prof.John Pavlopoulos .

Athens, Greece , October 2024

Dimitris Tsirmpas