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 moderation/facilitation is crucial for discussions to flourish and prevent polarization and toxicity, which nowdays seem omnipresent. However, being heavily based on humans, this moderation/facilitation proves costly, time-consuming and non-scalable, which has led many to turn to LLMs for discourse facilitation. In this thesis, we explore the use of LLMs as pseudo-users in online discussions, as a cost-efficient, realistic and scalable way of substituting initial LLM facilitation experiments, which would ordinarily necessitate costly human involvement. Furthermore, we show that including socio-demographic backgrounds in our LLM users leads to more realistic discussions. We explore the use of LLM annotators to estimate discussion quality, using a new statistical test to gauge annotator polarization, and 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]

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

[1]Code, datasets and analysis can be found at `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 αυτόματων επισημειωτών (με χρήση ΜΓΜ), για την αντιμετώπιση του προβλήματος της αξιολόγησης διαλόγων. Οι ψευτο-επισημειωτές αυτοί έχουν προκαθορισμένα από εμάς κοινωνικο-δημογραφικά υπόβαθρα, έτσι ώστε να προσομοιώσουμε τη διαφωνία που πιθανώς να υπάρχει ανάμεσα σε ανθρώπους με αντίστοιχα υπόβαθρα. Τέλος, αναλύουμε την επίδραση διαφόρων παραγόντων στην τοξικότητα των συνθετικών συζητήσεων, ως υποκατάστατη μετρική της ποιότητάς τους, χρησιμοποιώντας μεταξύ άλλων, έναν καινούριο στατιστικό έλεγχο για την εξήγηση πόλωσης μεταξύ επισημειωτών.

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 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 με ΜΓΜ.

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 Archimedes/Athena Research Center, whose input helped steer the project towards more productive directions. Lastly, I am grateful to my family, whose constant support has greatly aided me in all my endeavors.

Contents

1 Introduction

1.1 Motivation and Problem Statement

Large Language Models (LLMs) have revolutionized the field of Natural Language Processing (NLP) 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 [**ts2024**; **tan2024largelanguagemodelsdata**], have led to many established companies, startups and research groups around the globe to scramble and identify useful use-cases for this novel technology [**HadiASO**; **Zhou2024LargeLM**; **Hutchinson2024LLMAssistedVA**].

One such identified case is their use in online discussions. The online environment is essential to healthy democratic discourse [**WrightDemocracy**; **Janssen2005**; **Papacharissi2004DemocracyOC**] and deliberative discussions [**small2021polis**], whose goal is for citizens to share opinions in order to make informed decisions. However, because of the anonymity online discussions offer [**Avalle2024PersistentIP**], they are often characterized by aggression and toxicity [**XiaToxicity**], which often leads to low-quality discourse [**WrightDemocracy**] (although the latter position is contested [**Papacharissi2004DemocracyOC**]). Thus, discussions are often overseen by *"discourse moderators"*, people whose responsibility is to uphold the rules of the discussion and discipline users, and in more formal environments *"discourse facilitators"*, who ensure equal participation and help the participants coordinate. 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 [**Harvard2024**; **Wang2008StudentfacilitatorsRI**]. However, human facilitation is expensive, time-consuming and often relies on specialized staff [**small-polis-llm**].

LLMs are perfectly positioned to aid in facilitating discussions [**small-polis-llm**], since they are relatively inexpensive, can be scaled easily, and their summarization and text-generation abilities are ideal for the facilitation tasks we outlined above. However, finding the correct prompts and configurations (e.g. which model family, whether to use pretrained or finetuned models, ...) by use of robust experiments with human subjects can be 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 in 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 such as moderator strategy, moderator presence, and LLM user prompts. 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 research 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 standard facilitation tasks which are hypothesized to work with LLM facilitators, practical metrics for assessing argument quality, the risks and challenges of synthesizing discussions exclusively with LLMs, and existing datasets related to argument quality, synthetic discussions, and discourse facilitation.

Chapter 3

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

In this chapter, we describe the inner mechanisms of our framework. Section 3.1 (Requirements) outlines the functional and non-functional requirements for our new framework, explaining why existing frameworks fail to meet these needs. Section 3.2 (System Design) provides a high-level overview of the framework, detailing the synthetic discussion creation loop, the various user-configurable options, as well as the automated LLM annotation process. In Section 3.3 (Prompt Design), we discuss the different prompt templates and strategies used in both the synthetic creation and annotation processes. Lastly, Section 3.4 (Implementation) describes the framework's codebase, Application Programming Interface (API), and technical implementation details.

Chapter 4

This chapter details the experiments conducted in this thesis and their outcomes. In Section 4.1 (Experimental Setup) we describe the configurations and setup for the synthetic discussion creation and annotation tasks. Section 4.2 (Produced Datasets) presents the synthetic datasets generated by the framework during the experiments. Finally, in Section 4.3 (Results) we analyze the annotation results and examine how various factors impacted the quality (specifically, toxicity) of the synthetic 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 our experiments. Finally, we discuss the possible research avenues this thesis opens for future exploration.

Chapter 6

We briefly talk about the challenges and limitations of this thesis, both on a theoretical, and practical level. We also discuss the potential of our findings, and how we hope to build upon them.

2 Background and Related Work

This thesis intersects several scientific domains, including NLP, Social Sciences, Data Science, and Computer Science. Our focus lies particularly on discourse facilitation, argument quality, and conversational dynamics within Social Sciences, alongside the application of Large Language Models (LLMs) in synthetic dialogue generation, an area of interest within NLP. This section provides a focused review of these topics, detailing the rationale behind the selection of methodologies employed in the creation of our framework and subsequent analyses.

We first provide a basic background for human argumentation and the capabilities of LLMs (Section 2.1), and then dive into what has been achieved in the field of synthetic dialogue generation and discourse facilitation (Section 2.2).

2.1 Background

2.1.1 How and why do humans argue?

Collective deliberation and decision-making has been long hypothesized, and proven, to yield better results than those performed by individuals [**david-collaborative**; **stefan-dissent**]. 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. **graham2008disagree** proposes 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 [**dekock2022disagree**]. **walker-etal-2012-corpus** 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 [**dekock2022disagree**]. There are many attempts at refining the original hierarchy, such as the one presented by **benesch2016counterspeech**.

Disagreements and toxicity are a natural part of human dialogue, which however often lead to the discussion failing. **dekock2022disagree** 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. **Avalle2024PersistentIP** show that, from examining data of over the last 30 years, toxicity does not seem to discourage participation or escalate disagreements. Non-verbal discussions (newspaper comment sections, online discussions e.t.c.) nevertheless frequently cause participants to entrench themselves in their own beliefs, believing that the other participants are hostile to them, when exposed to toxic language.

2.1.2 The characteristics of online discussions

The above observation may lead us to conclude that online 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) [**boschi2021wordunderstandingsampleonline**]. 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 [**boschi2021wordunderstandingsampleonline**].

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, 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 [**Cornell_eRulemaking2017**].

Besides moderators, formal online discussion platforms and communities (as well as educational institutions [**Wang2008StudentfacilitatorsRI**; **Zhong2019ExploringTR**]) employ the use of discourse *facilitators*. While a moderator oversees discussions with the specific purpose of directing conversation towards predefined goals, a facilitator focuses on creating an environment that fosters collaboration and engagement among participants, thereby encouraging them to collectively achieve new insights or solutions [**wef_moderation**]. The differences between the two are subtle, and the terms are of-

ten used interchangeably, or as a combination of both duties in practice (such as in **Zhong2019ExploringTR**; **Carson2008**). We will thus use the term "facilitation" and "moderation" as to mean both responsibilities in this thesis.

2.1.3 What makes a good argument?

Both in popular perception and in academia, the best arguments are often considered to be the ones that sway public opinion, or that force the opposing side to concede previously held talking points. For instance, while the research of **zhang2016-oxford** 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.

We can not allow this notion of the best argument being the one that convinces the most people leak through when designing systems around deliberative and/or general online discussions. Should this happen, the culture of the platform will be one of highly-opinionated, heated discourse, between stubborn participants who try to "win" discussions by any means necessary. This phenomenon is mentioned by **karadzhov2023delidata**, who also point out that most existing datasets involve only two participants, whereas most online discussions and deliberative platforms usually involve groups of people interacting with each other.

2.1.4 Large Language Models

LLMs are sophisticated Artificial Intelligence (AI)-based computational models capable of text generation by training on vast amounts of written texts largely scrapped from the wider Internet. LLMs are based on the Transformer architecture [**vaswani2023attentionneed**], 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 [**ts2024**], as well as specialized annotation tasks [**tan2024largelanguagemodelsdata**]. 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 [**Vezhnevets2023GenerativeAM**; **aher2023usinglargelanguagemodels**] 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 [**Birkun_Gautam_2023**], 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

An area of intense LLM research is using LLMs to generate conversations among themselves. Researchers typically create a scenario where multiple agents discuss a given topic or a task, but instead of these agents being human, they are simulated by LLMs. The models are then finetuned on these conversations. Most approaches focus on strategies pitting a model against itself in an adversarial scenario [**liu2024largelanguagemodelsagents**; **cheng2024selfplayingadversariallanguagegame**; **zheng2024optimalllmalignmentsusing**], 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.

One of these RL adapted techniques is *"Self-play"*, where an agent learns by playing against itself rather than relying on a predefined set of opponents or scenarios. This method allows the agent to continually adapt and improve its strategies by facing progressively more challenging scenarios generated by its own evolving skills. Self-play has demonstrated spectacular results, outclassing human experts and rule-based computer algorithms in numerous games, the highest-profile being chess by the Alpha-Zero model in 2017 [**silver2017masteringchessshogiselfplay**]. Self-play can be applied to LLMs by making them talk to each other [**cheng2024selfplayingadversariallanguagegame**]. **ulmer2024bootstrappingllmbasedtaskorienteddialogue** propose a "Self-talk" framework where two LLMs are given roles ("client" and "agent") and a scenario which they act

out. The client is given a personality and freedom to choose its actions, while the agent is restrained to a few actions depending on the client's actions. Specifically, both are given a prompt containing their role, personality, and dialogue history. The client is provided with an intention, while the agent with appropriate instructions. The researchers demonstrate that self-talk can indeed be used to improve LLMs, given enough finetuning and rigorous filtering of input data. Notably for our thesis, it provides a practical demonstration that LLMs conversing with each other can produce quality conversations when applied in a structured setting, even if they are ultimately not used for model finetuning.

Moreover, LLM self-play is hypothesized to work in discourse facilitation tasks. **small-polis-llm** 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.

abdelnabi2024cooperationcompetitionmaliciousnessllmstakeholders 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 negotiation task itself is very difficult for most LLMs. We take inspiration from these experiments, and model one of our LLM prompting strategies after a competitive, scorable game.

It is important to note that conversations don't have to be constrained to only a few users. **park2022socialsimulacracreatingpopulated** show a novel technique of populating entire communities with hundreds of members with a technique called "Social Simulacra". This technique allows a single LLM instance to use a community's description, rules, and a set of a few dozens personality types, to populate a virtual community with posts and comments made by hundreds of users, having diverse personalities, goals and motivations. The researchers show that appropriately prompted LLMs using generated personas are adequate at mimicking human users, their posts being generally indistinguishable from the mirrored actual communities to human annotators. The idea of automatically generating personas to be used in synthetic dialogues can be very beneficial for frameworks aiming at generating them, such as the one presented in this thesis.

Finally, **lambert2024selfdirectedsyntheticdialoguesrevisions** follow the work of **Bai2022Constitution** and create a self-regulating conversation generation framework. Specifically, they use a set of given topics by **Castricato2024SuppressingPE**, which include various principles fundamentally based on human rights. They then define various conversation goals (e.g. help a user create an email, an essay, perform language translation etc.). An LLM then generates a plan (system prompt) for the conversation and begins generating the conversation according to that plan, while checking if at any point the principles have been violated. In that case, it generates a critique on why the 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 [**beck-etal-2024-sensitivity**]. An interesting specialized area where this technique is used is in LLM prompting ([**hwang-etal-2023-aligning**; **durmus2024measuringrepresentationsubjectiveglobal**] as cited by **beck-etal-2024-sensitivity**), where sociodemographic prompting can reduce misunderstandings between people belonging to different social groups by carefully phrasing its output.

beck-etal-2024-sensitivity demonstrate that incorporating sociodemographic information into LLM prompts can significantly enhance performance in various subjective NLP tasks under certain conditions. Large models (with over 11 billion parameters) often leverage combinations of sociodemographic traits rather than individual attributes, although they cannot treat these traits as explicit explanatory variables. However, this effect is highly dependent on factors such as prompt structure, model family, and model size, in ways that are not straightforward. Their findings thus support our hypothesis that incorporating user SDBs into LLM prompts can contribute to generating more diverse and realistic conversational outputs.

In addition to LLM sensitivities to sociodemographic prompts, the approach presents further limitations. Beyond the absence of standardized prompting templates and models capable of consistently leveraging sociodemographic information [**beck-etal-2024-sensitivity**], concerns have been raised regarding stereotypical biases [**cheng-etal-2023-marked**; **deshpande-etal-2023-toxicity**], as well as the strong orientation of models toward Western ideas and perspectives. There is also a lack of relevant datasets for languages other than English [**pmlr-v202-santurkar23a**; **durmus2024measuringrepresentationsubjectiveglobal**; **santy-etal-2023-nlpositionality**] as cited by **beck-etal-2024-sensitivity**. Additionally, **aher2023usinglargelanguagemodels** report the issue of sociodemographic "distortions,"

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

2.2.3 LLMs as discourse facilitators

LLMs are able to perform many facilitation tasks, which traditionally burdened human facilitators. One important use-case for LLMs is to iteratively summarize and refine the participants' understanding of the discussion and presented points. In a traditional discussion, a facilitator would present the participants with a summary of a key standpoint or worldview they presented as he understands it, and ask them whether the summary is correct [**small-polis-llm**; **Tsai2024Generative**]. This procedure continues iteratively until the group believes that the facilitator understands them. These points can later be used by the facilitators during intergroup discussion in order to test hypotheses about the different groups' opinions, which is especially useful in finding common ground. It is hypothesized [**small-polis-llm**] that using this procedure with an LLM may yield faster convergence to common ground and model understanding of the opinions of the participants.

small-polis-llm further point to using LLMs to directly produce opinions at the start of the dialogue (called "seed opinions" in the original paper) as another area of interest. However, **karadzhov2023delidata** demonstrate that synthetic data 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.

An important part of facilitation is identifying the best course of action in various emergent situations. **al-khatib-etal-2018-modeling** analyze a deliberative discussion in terms of "deliberative strategies", which are comprised of a sequence of "moves" each participant can take during the discussion. We hypothesize that a LLM facilitator could look at the current state of discussion and recommend the best possible move according to the best possible strategy to the participant. It is worth noting that the researchers define the goal of a deliberative discussion differently than the one used by this paper and defined by Polis [**small-polis-llm**]. 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-2021-towards approach the problem more directly, reporting on human moderators and how their behavior should be modeled by automated systems. They provide an example where a moderator handles two users with different positions and argument

styles who were in the process of derailing the discussion, and another where a user directly confronts the moderator on the definition of the forum's rules. Human moderators typically follow standard guidelines on how to approach situations such as these, as well as how to facilitate discussion, as discussed above. Thus, synthetic moderators should be modeled after these interactions and guidelines.

Finally, we note that LLMs are well positioned to tackle traditional NLP problems relating to facilitating online discussions; namely machine translation (in order to allow marginalized and minority groups to contribute) [**Tsai2024Generative**], hate-speech [**Nirmal2024TowardsIH**; **shi-2024-hatespeech**], toxicity [**kang-qian-2024-implanting**; **Wang2022ToxicityDW**] and fake-news [**Liu2024DetectIJ**; **Xu2024ACS**] detection, in order to ensure effective moderation.

2.2.4 Measuring Discussion Quality

vecchi-2021-towards 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) [**Steiner2005-STEDPI-8**; **stab-gurevych-2017-parsing**], 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 [**wachsmuth-etal-2017-computational**].

dekock2022disagree 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 LLM *moderator's* performance on actual conversations, we still lack a way of quantifying the quality/realism of the *synthetic* dialogues. **ulmer2024bootstrappingllmbasedtaskorienteddialogue** propose a series of automated evaluation metrics for synthetic dialogues. Non-task-specific metric include "Dialogue Diversity" which counts the number of n-grams (unigrams up to 5-grams) and the pairwise ROUGE-L [**lin-2004-rouge**] score between the outputs of a LLM in a single interaction and "Character Consistency", which measures how much the LLM stays "in-character" and which is evaluated by a finetuned DeBERTa [**he2023debertav3improvingdebertausing**] model. The researchers show a strong correlation between these metrics and subjective discussion quality evaluations.

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

2.2.5 Risks and Challenges

First, we feel compelled to echo the warnings of **small-polis-llm** that 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 **park2022socialsimulacracreatingpopulated**.

small-polis-llm additionally outline several known weak points in LLM usage for moderation/facilitation; LLMs suffer from bias, hallucinations, are vulnerable to prompt injection attacks, and have their own political leanings (with most trending towards progressive ideas [**Taubenfeld2024SystematicBI**]). Furthermore, **vecchi-2021-towards** note that care must also be taken when quantifying argument quality by measures such as likes, to ensure the model does not discriminate against users who do not belong to a prevalent group or have difficulty communicating, as would be the case in frameworks such as Polis [**small2021polis**]. They also recommend using discussions from online message boards for the initial synthetic comments ("seed opinions"). **vecchi-2021-towards** however, warn of the challenges of sourcing such comments, since personal opinions, facts and fake news are often bundled together in online discussions.

Lastly, training generative models, and more specifically LLMs, on their own data most often leads to the model collapsing [**alemohammad2023selfconsuminggenerativemodelsmad**; **shumailov2024curserecursiontraininggenerated**] as cited by **ulmer2024bootstrappingllmbasedtaskoriented**
Even when not trained on their own data, LLMs tasked with creating dialogues often generate low quality, off-topic and generally useless conversations.In their experiments, **ulmer2024bootstrappingllmbasedtaskorienteddialogue** show that, at many points, the conversation collapses. Their actors in this case go off-script, begin rambling or end the interaction too early or too late. Other challenges include hard and soft errors when generating data at-scale [**lambert2024selfdirectedsyntheticdialoguesrevisions**; **ulmer2024bootstrappingllmbasedtaskorienteddialogue**] requiring automatic verification steps, insidious errors which can not be reasonably caught by automated metrics [**lambert2024selfdirectedsyntheticdialoguesrevisions**; **ulmer2024bootstrappingllmbasedtaskorienteddialo**
and a lack of generated topic diversity [**lambert2024selfdirectedsyntheticdialoguesrevisions**].

2.2.6 Related Datasets

Synthetic-only dialogue datasets are exceedingly rare in literature. [**lambert2024selfdirectedsyntheticdial**
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. Thus, we need to explore datasets from adjacent tasks to aid us in analyzing and evaluating our own discussions in the future.

One of the most frequently used datasets for conversation escalation analysis is the "Wikipedia Disputes" dataset [**de-kock-vlachos-2021-beg**], which contains discussions from Wikipedia's talk pages, where members attempt to resolve edit disputes. The annotated labels correspond to whether a dispute "escalated", meaning that the members could not resolve it by themselves, and thus requested moderator arbitration. **dekock2022disagree** build upon their work and provide the "WikiTactics" dataset, which provides annotations based on the tactics employed in each comment. **hua2018wikiconvcorpuscompleteconversational** enhance the "Wikipedia Disputes" dataset, including metadata concerning edits, deletions and other actions on the comments themselves. This approach was followed by **al-khatib-etal-2018-modeling** who provide a large-scale dataset generated from Wikipedia discussions, called "Webis-WikiDebate-18 corpus", designed to model deliberative discourse based on metadata categories. The dataset contains $2,400$ turns labeled with discourse acts, $7,437$ turns labeled with relational connections between comments, and $182,321$ turns labeled with discourse frames. Each turn in the discussion is labeled automatically using metadata that corresponds to specific discourse categories derived from their own discourse classification models.

Early conversation derailment datasets are also available, albeit in relatively small numbers. These datasets are useful for diagnosing the causes of conversational collapse in human dialogues. **zhang-2018-gone-awry** provide a curated dataset of $1,270$ conversations with an average length of 4.6 comments each, featuring derailed conversations. **chang-danescu-niculescu-mizil-2019-trouble** provide two datasets relating to discussion derailment, the first expanding on the previous dataset with a total size of $4,188$ conversations and a larger discussion length, while the second is sourced from the "Change My View" (CMV) Subreddit, featuring $600,000$ conversations, $6,842$ of which necessitated moderator intervention.

3 System Design and Implementation

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

3.1 Requirements

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 [**Vezhnevets2023GenerativeAM**] and LangChain [**langchain**] fit, or can be made to fit, all functional requirements listed above. They however fail in almost all non-functional requirements. A major issue is that their APIs are convoluted, although this could be circumvented with enough effort by employing the Adapter pattern [**gamma1995design**]. Another more serious problem, however, is 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. Given that this thesis was developed under acute resource constraints, the latter issue could not be easily resolved.

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

3.2 System Design

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

3.2.1 Synthetic Dialogue Creation

We use a simplified version of the LLM discussion framework outlined in **abdelnabi2024cooperationcompet**
Each actor is given his turn to speak according to a round-robin scheduling algorithm

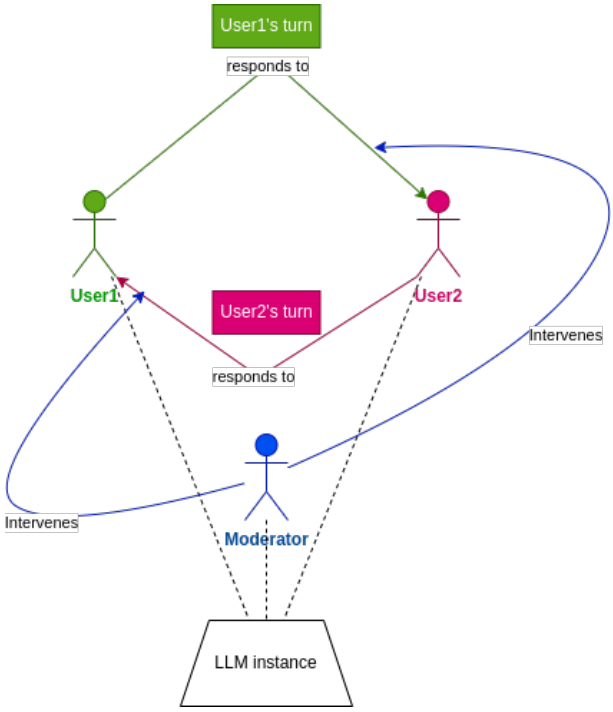

Fig. 3.1: The conversation loop on which the SDF operates. Can be generalized for N users and 0 or 1 moderators.

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

The users and the moderator are all controlled by the same LLM instance (like in **park2022socialsimulacracreatingpo** we only change the system prompt when each takes their turn to speak. The prompts are comprised of five parts:

- **Name**, the name of the actor, used for other actors to refer to him in-conversation.
- **Role**, the role of the actor within the conversation (user or 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.

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])

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

3.2.2 Automated Dialogue Annotation

As per the non-functional requirements of Section 3.1, we need a mechanism which can automatically annotate already-executed conversations. This could be achieved by using specialized classification models such as a model for toxicity classification, another for argument quality, and so on. However, these usually differ not only on their exact architecture, but also on their fundamental type; for instance, in toxicity classification, competitive models can be Machine Learning (ML)-based instead of Deep Learning (DL)-based [**anjum2024hate**]. 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 reliably 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.

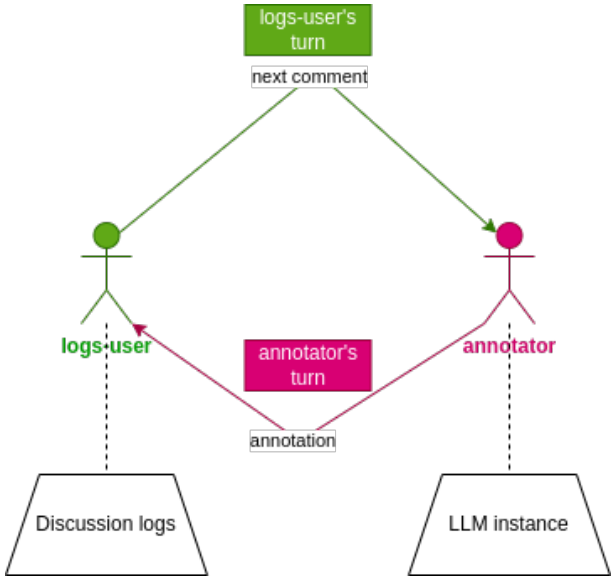

Fig. 3.2: The annotation loop on which the SDF operates. Note the purposeful similarity of the function to Figure 3.1.

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

In this thesis, we use the second option.

We re-use the 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 (h) for the annotator; in other words, for each comment, the annotator can see the h preceding comments of the conversation, in order to ensure he understands the context under which each comment was made. 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

In order to create convincing prompts for both facilitators and users, we first had to define the rules of the discussion, which necessitated the definition of a policy for our virtual chat-room. We defined this policy according to how an ideal online/deliberative discussion would look like.

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

Our facilitator prompts were largely designed on preventing these two phenomena. We also utilize some general guidelines solicited from the Cornell eRulemaking moderator manual [**Cornell_eRulemaking2017**]:

- 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.

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

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

- Many people are unwilling to change opinions during online discussions.
- Personal attacks are common [**dekock2022disagree**].

- Anonymity makes people more likely to assume other participants are not arguing in good faith, especially in toxic environments [**Avalle2024PersistentIP**].

- People are not told what to believe when entering a discussion, but form opinions based on (besides many other factors) their SDB.

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

3.3.2 "Moderation Game" prompts

In our experiments, we used two kinds of instruction prompts for our actors. One is a "standard" instruction prompt 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 **abdelnabi2024cooperationcompetitionmaliciousnessllmstakeholders**. 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

The assigned actions and weights were somewhat arbitrary. In the future, we hope to use RL or some selection strategy in order to tune them in a way that is optimal for facilitation, according to synthetic experiments.

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 [**gamma1995design**] 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 [**gamma1995design**], containing Model Adapter classes from **models.py** and providing 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.) 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").

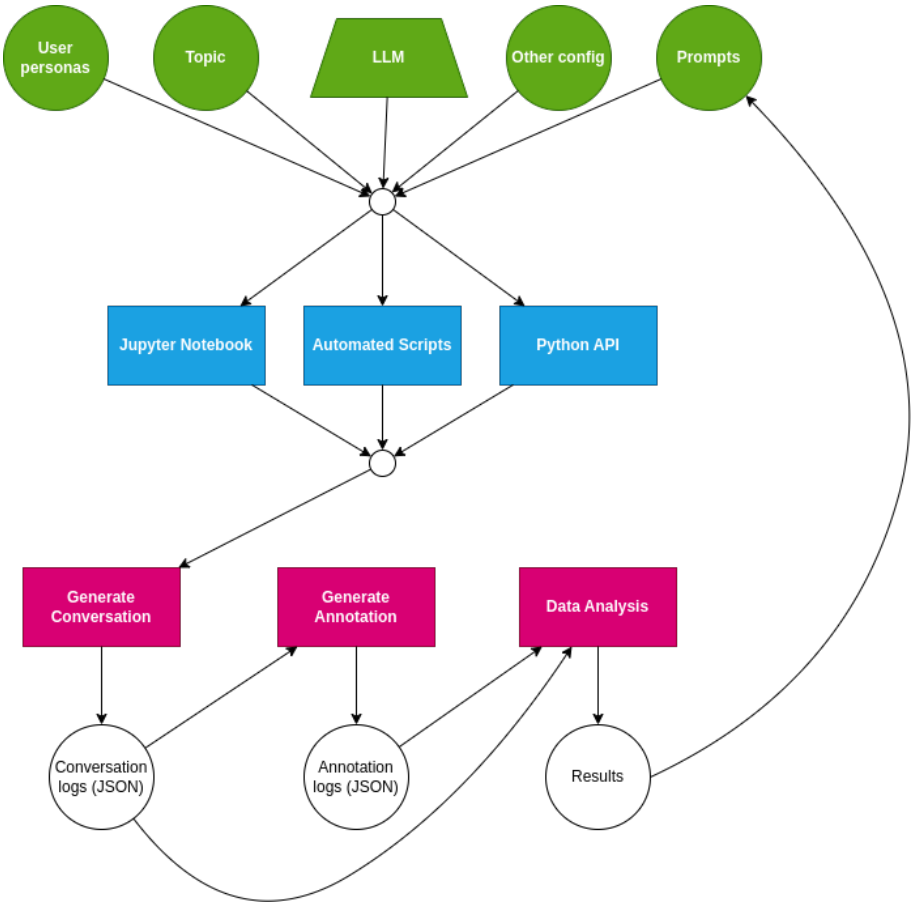

Fig. 3.3: An abstract view of the SDF. Green shapes represent various configurations, blue shapes entry points (see Section 3.4.2), pink ones processes delegated to the SDL, and **white** ones exported data.

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 4.2. The topics were adapted from **pavlopoulos-likas-2024-pol** and in some cases abbreviated in order to fit the input context limits of the local LLM instance.

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

3. Select whether a moderator is present.

4. Provide actors with the global context.

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

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

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

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

The combination of the settings above produces a set of JSON configuration files, each containing the configuration for a single experiment. We execute all possible combinations, except for "freezing" conversation-specific configurations. As such, we use a window of 4 comments and a conversation length of 3 round-robin rounds. It is worth noting that we

also avoid pitting the African American SDB against the LGBTQ+ SDB, in order to ensure that the discussion will be polarized.

Tab. 4.2: Controversial topics used as seeds for the simulated conversations. Excerpts selected from **pavlopoulos-likas-2024-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 someone a cake because it is for a gay wedding is no different from selling someone a cake because it is for a mixed race wedding or a Muslim wedding or any other kind of wedding. IT IS DISCRIMINATION! Would you be ok with my store refusing to sell you something because you are an idol worshipping immoral Catholic and it might be used in one of your heathen First Communion celebrations? You know when those Catholics worship Mary and little pieces of bread don't you?
5	How do you know he is not proud of his cakes? Artists do not take pride in their work? Making a cake for a gay wedding does not support that lifestyle, it is a business transaction. Period. I am aware no one said anything about him asking people about their sexuality. I am sorry that was hard for you to understand. Is he going to ask everyone that comes in if the cake is for a gay wedding? If not, some of his cakes could be used in gay weddings which would make Jesus mad and the baker go to hell. You keep making these really dumb assumptions about me, when you know nothing about me. I am not confused, you are rude. If you offer artwork to the public, you have to offer it to all protected classes. Why would black people be discriminated against? Precedent. Ridiculous? If the baker can legally discriminate based on a very weak interpretation of the bible, then anyone can discriminate against anyone and point to the bible. Satanists can discriminate against Christians...

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

4.1.2 Automated Dialogue Annotation

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

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

For the purposes of analyzing inter-annotator agreement, we use the normalized Distance From Unimodality (nDFU) [**pavlopoulos-likas-2024-polarized**], a measure used to evaluate the polarization of ratings (e.g. text annotations using a Likert scale) in a data set. DFU, or Distance From Unimodality, is originally defined as the maximum difference of the frequency bins in a histogram, when descending sidewise from (from the right or the left of) the mode. To normalize DFU, it is divided by the mode, in order for the measure to be within the $[0, 1]$ range, instead of DFU's $[0, freq_{mode}]$.

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

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

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

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

4.2 Produced Datasets

We produce three synthetic datasets:

- 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 [**pavlopoulos-likas-2024-polarized**] score of each comment, and the individual annotations for each annotator SDB.

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

4.3 Results

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.

4.3.1 Observations on the behavior of synthetic user SDBs

In this section, we investigate the following hypothesis: **LLM chat users change their behavior according to the supplied SDB**.

Table 4.5 presents the expected and actual behavior of synthetic users from different SDBs across the two polarized topics. The seed comments used to start the synthetic conversations were split into two political subjects; racism on the basis of racial identity, and on the basis of sexual orientation. On the one hand, it's clear that SDBs significantly alter the behavior of our synthetic users. On the other hand, our 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 African American users, who were always the first to speak (and thus could not be influenced by the stances of other users), always adopted a progressive stance in topics concerning the LGBTQ+. This is not a behavior which necessarily matches with reality, since, in the real world, many African Americans may adopt conservative stances [**lockerbie2013race**; **mckenzie2013shades**]. We hypothesize that this discrepancy is caused by the LLM's societal and political biases; the model may have assumed that, since African-Americans are a minority, they should also be progressive in other social issues. In contrast to the other SDBs, 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 stances influenced from the model's own biases (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 topics and configurations**. 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" [1] (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 (statistically and qualitatively) 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.

[1] See caveat on whether African Americans can be considered progressive in Section 4.3.1

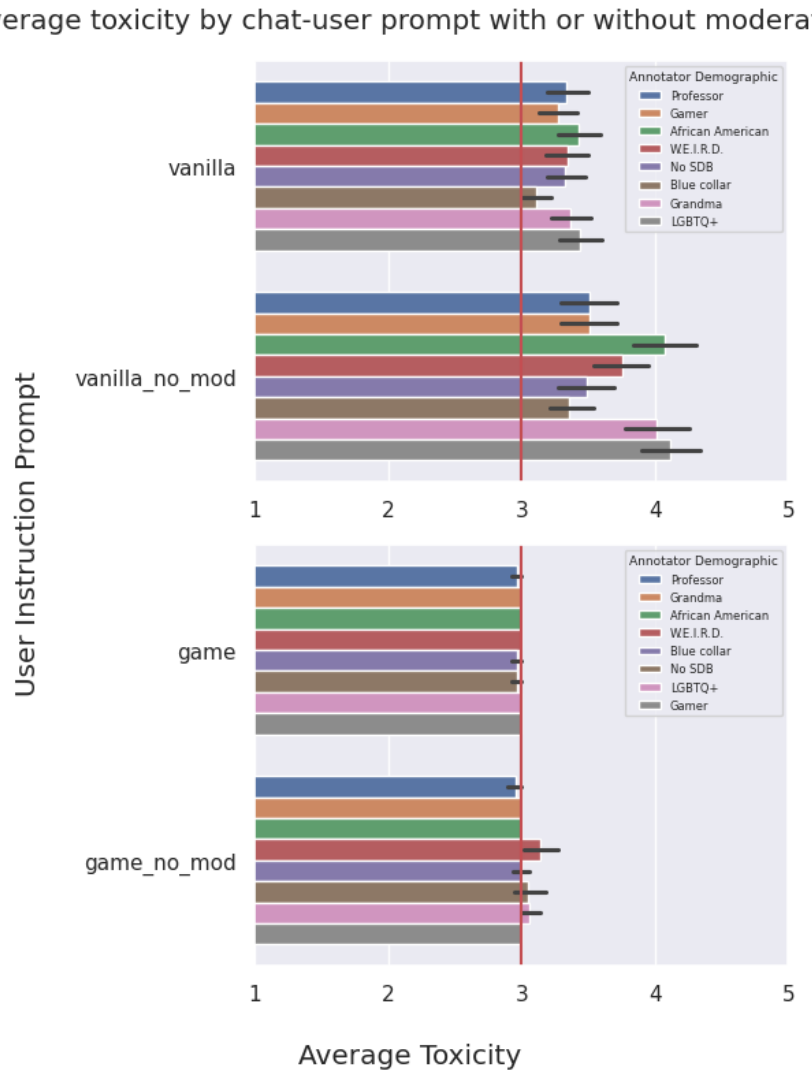

Fig. 4.1: Mean toxicity by prompting strategy and moderator presence, per annotator SDB.

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.*

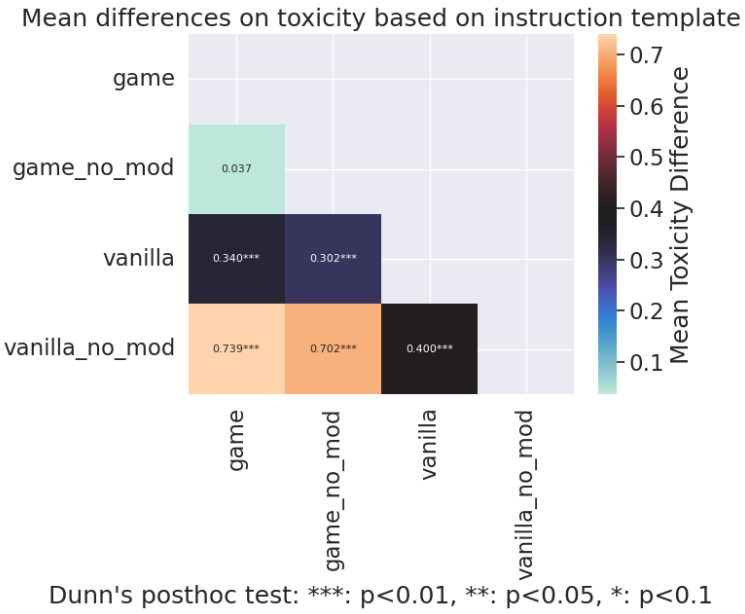

Fig. 4.2: Mean annotation difference between each strategy/moderator presence. Each comparison is accompanied by Dunn's posthoc test for multiple comparisons in the form of significance asterisks.

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 first check whether disagreement exists between the various annotations. Figure 4.3 shows the nDFU[**pavlopoulos-likas-2024-polarized**] scores for each synthetically created comment. The majority of comments are in perfect annotator agreement ($nDFU$=0), while a few are in perfect disagreement ($nDFU$=1).

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

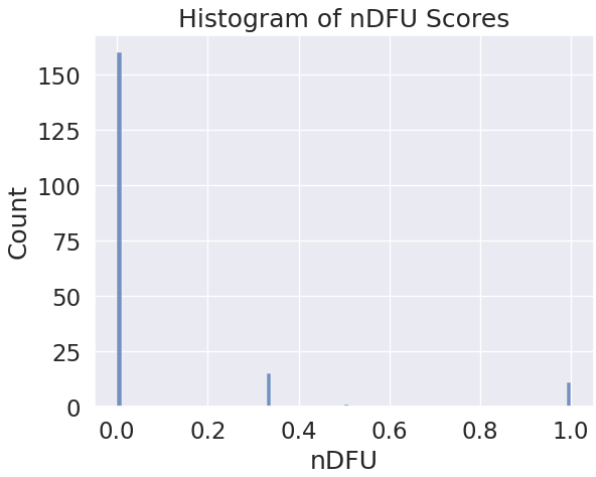

Fig. 4.3: nDFU [**pavlopoulos-likas-2024-polarized**] scores for each comment. More is larger disagreement between the annotators.

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.

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[2], 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, 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. Note that our expected behavioral model is completely different from the actual annotations, indicating that SDBs have failed to model human annotators.

[2]See caveat on whether African Americans can be considered progressive in Section 4.3.1

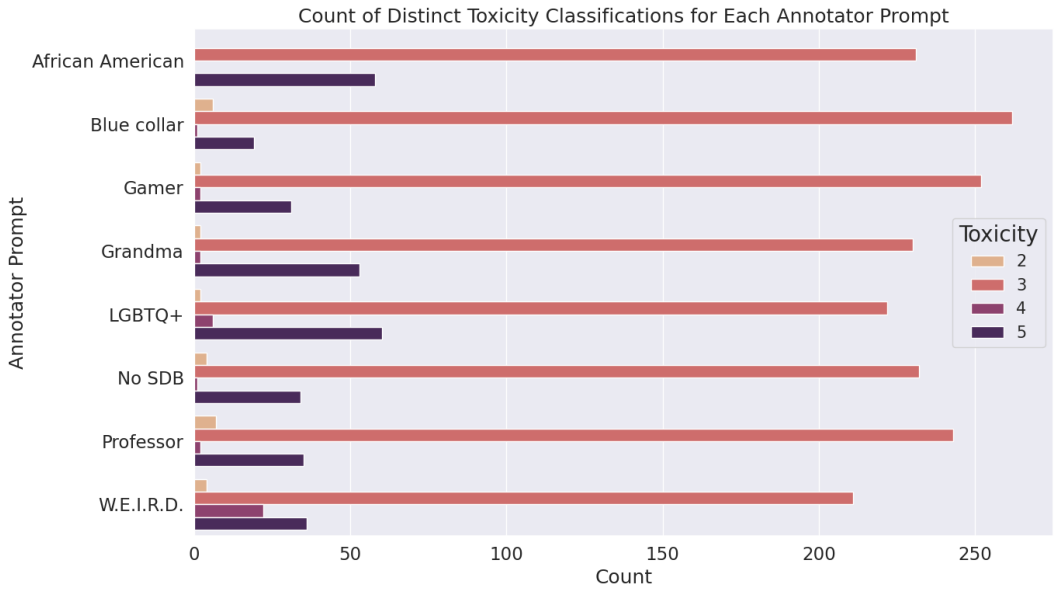

Fig. 4.4: Toxicity annotations by annotator SDB prompt. Note the high preference towards group 3 ("moderately toxic") and that significant deviations only occur between groups 4 ("very toxic") and 5 ("extremely toxic").

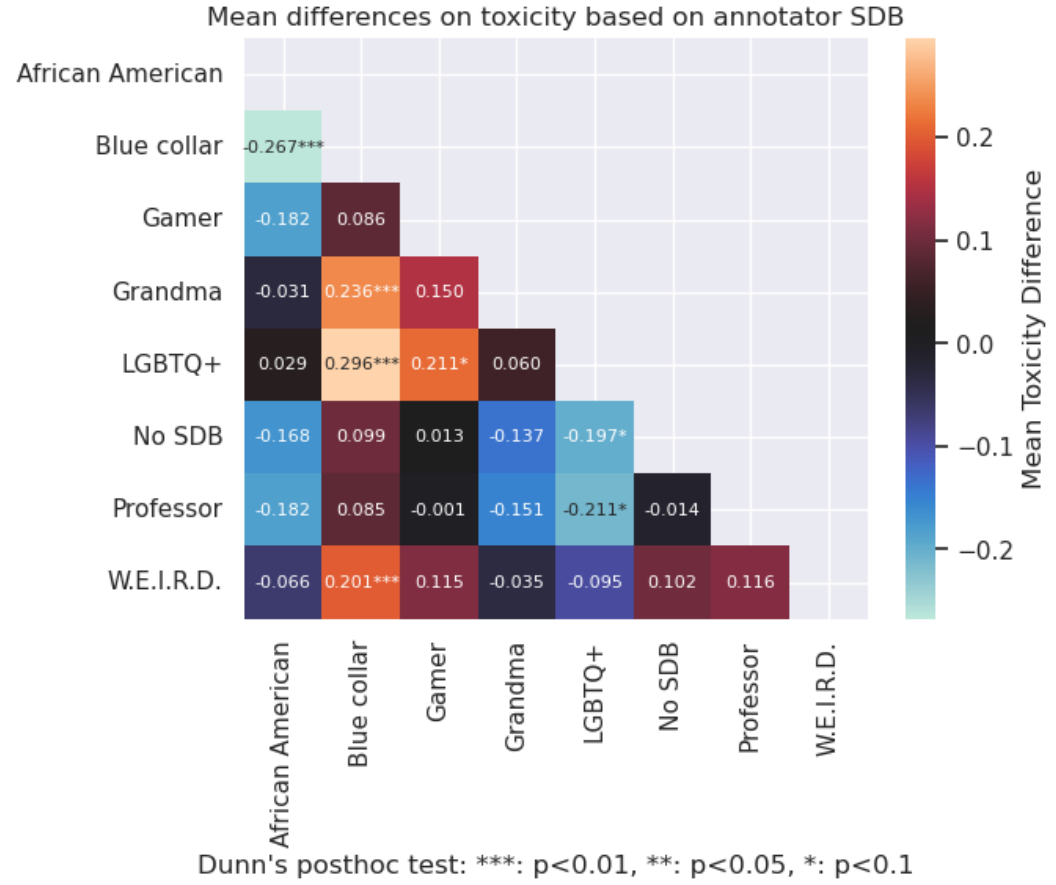

Fig. 4.5: Mean annotation difference between each annotator SDB. Each comparison is accompanied by Dunn's posthoc test for multiple comparisons in the form of significance asterisks.

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-likas-2024-polarized**. This measure compares the nDFU of the set comprising all the annotations, with the nDFUs of each individual annotation set, partitioned by the factors of a selected feature. In mathematical terms, let X the set with all annotations and $X_i, i \in G$ the set comprising all annotations where the annotator has characteristic i, and G the set of all characteristics (factors) within a feature. Then, feature G explains the polarization in X if $nDFU(X)\ \ 0$, but $nDFU(X_i)\ \ 0, \forall i \in G$.

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".

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

A Wilcoxon signed-rank test is then performed to assess whether the nDFUs of all factors are statistically indistinguishable from zero (e.g. $nDFU(X_{men})$=$nDFU(X_{women})$=0). Since annotation data typically do not follow a normal distribution and are often limited in number, the non-parametric Wilcoxon test is chosen for its robustness [**Rey2011**]. The null hypothesis (H_0) posits that the feature does not explain the observed polarization (i.e., all nDFUs are zero), and since $nDFU(a) \in [0, 1] \forall a$, the alternative hypothesis (H_a) is that $\exists i \in G : nDFU(X_i)\ \ 0$. Finally, the algorithm outputs both the global nDFU and the complement of the p-value ($1 - p$), where a low value indicates strong evidence against aposteriori unimodality, suggesting that the feature likely contributes to the observed polarization. In our example, we can not claim that *gender* influences annotation polarization if $nDFU(X_{men})$>0 or $nDFU(X_{women})$> or $nDFU_{global} \approx 0$.

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

robust test would check whether the nDFUs of the individual factors are statistically smaller than the global (e.g. $nDFU(X_{men})$<$nDFU(X_{global})$ or $nDFU(X_{women})$<$nDFU(X_{global})$.

Algorithm 3 Our proposed Aposteriori Unimodality Test

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

```
all_annotations ← concatenate(grouped_annotations_by_factor)
global_ndfu ← ndfu(all_annotations)

ndfus ← {}
for group in grouped_annotations_by_factor do
    ndfus ← ndfus ∪ ndfu(group)

p ← wilcoxon(ndfus, 0, alternative "greater")
return (global_ndfu, 1 − p)
```

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

	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.

5 Conclusions & Future Work

In this thesis, we explored 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; standard 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 our synthetic dataset, we found that the presence of a moderator/facilitator can be a decisive influence on the toxicity of a discussion. Furthermore, framing the discussion as a scorable game seems to potentially keep LLM users in line using the threat of a moderator whose presence may not be perceivable. Finally, we defined a new statistical test that attributes polarization to specific SDB features. By using this test alongside many other techniques we can not decisively prove that using different SDBs in LLM annotators yields no significant qualitative difference, and that any difference could be attributed to a change in priors, as opposed to the annotator 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 moderation/facilitation techniques, interventions and LLM families 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.

6 Discussion

The initial goal of this thesis was to develop a framework for generating synthetic dialogues to support research into automated moderation/facilitation techniques. As the project evolved, a key concern arose: whether the synthetic dialogue setup was truly representative of human interaction. This prompted us to conduct our own experiments to explore this issue in more detail.

Our findings were promising. We observed that interventions by LLM moderators had a notable impact on reducing toxicity within synthetic discussions. Additionally, we found that LLMs, when provided with only a SDB prompt, were able to convincingly adopt and maintain specific positions in the conversations. This provides us with a concrete incentive to continue development, and experimentation on, synthetic dialogues.

However, several inconsistencies were also observed in the behavior of both the LLM-generated users and the annotators involved in the experiments. These inconsistencies could be traced back to multiple factors, including non-optimal prompt design, potential bias in the models, and limitations inherent in the models themselves. A significant constraint in our research was the use of a smaller and now outdated LLM, driven by resource limitations. This constrained both the quality of the synthetic conversations and the input context width available to us.

Additionally, our need for automated annotation led us to explore ways with which to gauge and attribute polarization among different annotator features. While promising, our proposed statistical test can be significantly improved by modifying the null hypothesis (H_0).

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

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