ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS

School of Information Sciences and Technology
Department of Informatics
Athens, Greece

Master Thesis
in
Computer Science

Mitigating Polarisation in Online Discussions Through Adaptive Moderation Techniques

Dimitris Tsirmpas

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

Committee: Assoc. Prof. Ioannis Pavlopoulos
Department of Informatics
Athens University of Economics and Business

Asst. Prof. Richard Miles
Department of Informatics and Telecommunications
University of Athens

June 2024

Dimitris Tsirmpas

Mitigating Polarisation in Online Discussions Through Adaptive Moderation Techniques

June 2024

Supervisor: Prof. Ion Androutsopoulos

Athens University of Economics and Business

School of Information Sciences and Technology

Department of Informatics

Mobile Multimedia Laboratory

Athens, Greece

Abstract

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

This is the second paragraph. Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

And after the second paragraph follows the third paragraph. Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

After this fourth paragraph, we start a new paragraph sequence. Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are

written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

This is the second paragraph. Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

Περίληψη

Εδώ γράφουμε μια περίληψη της διπλωματικής μας στα Ελληνικά. Πρέπει να περιλαμβάνει τουλάχιστον το περιεχόμενο του Αγγλικού abstract, αλλά καλό είναι να έχει λίγο μεγαλύτερη έκταση και να δίνει μια πλήρη (αλλά σύντομη!) εικόνα του τι πραγματεύεται η εργασία.

Acknowledgements

And after the second paragraph follows the third paragraph. Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language. Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

After this fourth paragraph, we start a new paragraph sequence. Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language. Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language. Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

Contents

1 Introduction

This is the second paragraph. Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

And after the second paragraph follows the third paragraph. Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

1.1 Motivation and Problem Statement

After this fourth paragraph, we start a new paragraph sequence. Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected

font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

1.2 Thesis Structure

Chapter 2

This is the second paragraph. Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

Chapter 3

And after the second paragraph follows the third paragraph. Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

Chapter 4

After this fourth paragraph, we start a new paragraph sequence. Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

Chapter 5

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest

	Name	Task	Type
1	Tom	Seek	Cat
2	Jerry	Hide	Mouse

Tab. 1.1: This is a test table

gefburn"? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

Table 1.1 refers to the results in Figure 4.1.

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 the individual [**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**] 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 [**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 [**benesch2016counterspeech**].

Disagreements and toxicity are a natural part of human dialogue, which 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 it will be issued both by the same person and/or by another participant in the future, and leading to discussions failing on average. 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 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 believes, believing that the other participants are hostile to them, when exposed to toxic language.

It is worth noting that online conversations differ greatly from offline (face-to-face) conversations. Online forums are typically larger in terms of lengths and participants, forming large trees of replies to 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, facilitation of effective communication, and addressing any issue that may arise during the course of the proceedings. In informal communities, moderators are respected members of the community, while in more formal settings, may be paid employees. In both cases, but especially the latter, moderators are given a set of special rules and guidelines they are 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**].

2.1.2 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 opposite side to concede previously held talking points. For instance, while the research of [**zhang2016-oxford**] claims to investigate how ideas flow between groups holding and discussing different views, and while its insights are doubtlessly important, ultimately misses its stated goal by subscribing to this notion. The authors end up investigating what wins an argument, their analysis quickly pivoting to audience reactions, votes, rhetorical dominance and predictive modeling for which team is likely to win a debate, instead of how ideas influence the discourse itself.

In a system which aims 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 [**karadzhov2023delidata**], alongside the fact that most existing datasets involve only two participants, whereas deliberating platforms usually involve group thinking and deliberation.

Another approach to ranking argument quality is through a combination of linguistic metrics such as grammar quality, clarity or degree of relavance and overall impact to the discussion [**Gretz2019ALD**].

2.1.3 Large Language Models

Large Language Models (LLMs) are sophisticated artificial intelligence systems designed to understand and generate human-like text by processing vast amounts of language data. LLMs are based on the Transformers architecture [**vaswani2023attentionneed**], after it was widely adopted in numerous models undertaking many Natural Language Processing 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, as well as specialized annotation tasks. 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 e.t.c.), 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 humans 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 redudant.

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 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 utilizing the model talking with itself has become an area of intense research into LLMs. 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 Reinforcement Learning. However, not all self-talking approaches use Reinforcement Learning or an adversarial scenario, nor are they used exclusively in the context of jailbreak prevention.

[**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 (scores attached), 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 Reinforcement Learning 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 [**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. 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 [**MosaicML2023**] model for the client, and a 7 billion

parameter same model 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.

[**lambert2024selfdirectedsyntheticdialoguesrevisions**] follow the work of [**Bai2022ConstitutionalAH**] and create a self-regulating conversation generation framework. Specifically, they use a set of given topics by [**Castricato2024SuppressingPE**] and define the conversation goals. The LLM then generates a plan (system prompt) for the conversation with itself, checks if at any point the goals have been violated, and if so 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. However the study failed to find any trends between goals and the generated text.

2.2.2 LLMs bearing sociodemographic background

Including a sociodemographic (SD) background (race, age, ethnicity e.t.c.) 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**; **durmus2024measuringrepresentationsubjectiv**] 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 including sociodemographic information in LLM prompts can in some situations greatly increase their performance in various NLP tasks. More specifically, they show that changing sociodemographic information significantly influences classification results (which is also observed between humans of different social and demographical groups), although the results are contingent on the prompt structure, model family and model size, in non-obvious ways. Large models (containing more than 11B 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 [**beck-etal-2024-sensitivity**], skepticism exists concerning stereotypical biases [**cheng-etal-2023-marked**; **deshpande-etal-2023-toxicity**] as well as models

having a large bias towards responses from Western countries, and the unavailability of relevant datasets concerning languages other than English [**pmlr-v202-santurkar23a**; **durmus2024measuringrepresentationsubjectiveglobal**; **santy-etal-2023-nlpositionality**] as cited by [**beck-etal-2024-sensitivity**]. Furthermore, [**aher2023usinglargelanguagemodels**] 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 SD 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 aluminium.

Finally, we need to voice our own practical and ethical concerns on sociodemographic prompting. While undoubtedly a useful technique, especially when it comes to synthetic data creation, gathering, storing, and feeding personal user data into AI models may be in violation of both GDPR [**gdpr**] and the EU AI Act [**eu_ai_act_2021**]. This may lead to situations where only a subset of the sociodemographic information, on which the model was trained and evaluated on, can be used in a practical application.

2.2.3 LLMs as discourse facilitators

[**small-polis-llm**] deliberate and experiment on ways LLMs can be applied to Polis [**small2021polis**], a discussion and deliberation platform which aims to facilitate constructive dialogue and collective decision-making by modeling public opinion through machine learning and human interaction. It allows participants to submit and vote on comments, using techniques like Principal Component Analysis and K-means clustering to visualize and identify consensus and distinct opinion groups. Participants can interact with these groups by reading, voting on, and adding their own comments on the discussion, which are then curated by facilitators, who are also active members of the discussion. As of writing, Polis does not leverage any Natural Language Processing (NLP) methods. The authors propose using LLMs to automate various tasks previously necessitating direct human involvement. For the purposes of this thesis, we only consider points related to discussion participation and moderation.

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. 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. The authors hypothesize 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 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. The authors additionally 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, which further solidifies the theoretical base of this thesis. However, [**karadzhov2023delidata**] 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-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. 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 [**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**] 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 facilitating discussion, as discussed above. Thus, automated moderators should be modelled after these interactions and guidelines, as well as more traditional hate-speech, fake-news and trolling detection, to ensure effective moderation.

2.2.4 Measuring Argument 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 consider-

ation 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 LLMs performance on actual conversations, we still lack a way of quantifying the quality of the synthetic dialogues. [**ulmer2024bootstrappingllmbasedtaskorienteddialogue**] 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 [**lin-2004-rouge**] 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 [**urbanek-etal-2019-learning**] 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 [**he2023debertav3improvingdebertausing**] model.

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. 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.

2.2.5 Risks and Challenges

Firstly, we feel compelled to echo the author's warnings in [**small-polis-llm**]. Synthetic data and conversations should by no means replace human content and interactions. This thesis builds a theoretical base for future models, trained and deployed on human-to-human discussions (with the presence of LLM actors) and monitored by human moderators. A more pressing concern would be the use of this research on the development of social-network troll/bot farms, as also expressed by [**park2022socialsimulacracreatingpopulated**].

[**small-polis-llm**] 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-2021-towards**] 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 [**small2021polis**].

Lastly, training generative models, and more specifically LLMs, on their own data most often leads to the model collapsing [**alemohammad2023selfconsuminggenerativemodelsmad**; **shumailov2024curserecursiontrainginggenerated**] as cited by [**ulmer2024bootstrappingllmbasedtasl**
Additionally, even when not trained on their own data, LLMs tasked with creating dialogues often generate low quality, off-topic and generally useless data [**ulmer2024bootstrappingllmbasedtaskori**
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 [**lambert2024selfdirectedsyntheticdialoguesrevi**
ulmer2024bootstrappingllmbasedtaskorienteddialogue] requiring automatic verification steps, insidious errors which can not be reasonably caught by automated metrics [**lambert2024selfdirectedsyntheticdialoguesrevisions**; **ulmer2024bootstrappingllmbasedtaskorien**
and generated topic diversity [**lambert2024selfdirectedsyntheticdialoguesrevisions**].

2.2.6 Datasets

[**small-polis-llm**] recommend using discussions from online message boards for the initial synthetic comments ("seed opinions"). [**vecchi-2021-towards**] however, warn the challenges of sourcing such comments; personal opinions, facts and fake news are often bundled together and the language used in many social media platforms is significantly different from the one used in deliberation platforms (such as Polis).

One of the most frequently used datasets for goal-oriented discussions is the Wikipedia Disputes dataset [**de-kock-vlachos-2021-beg**], which contains discussion on the Wikipedia's talk pages, where members attempt to resolve edit disputes. The annotated labels corre-

spond to whether a dispute got "escalated", meaning that the members could not resolve it by themselves and requested moderator arbitration. [**dekock2022disagree**] 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. [**hua2018wikiconvcorpuscompleteconversational**] 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. [**al-khatib-etal-2018-modeling**] enhances the WikiDebate dataset by including metadata on deliberative strategies employed by each user.

Early convresation derailment datasets are also available, albeit in relatively small numbers. [**zhang-2018-gone-awry**] provide a curated dataset of 1270 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, 6842 of which necessitated moderator intervention.

One of the few datasets containing group discussions is the "Deli Data-Deliberation Dataset" [**karadzhov2023delidata**], which includes 500 group discussions, and is annotated by both metadata and an objective measure of decision correctness. 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 for 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 provided by [**lambert2024selfdirectedsyntheticdialoguesrevisions**], ...

[**zhang2016-oxford**] provide a dataset from the "Intelligence Squared" (IQ2) Oxford debates. Their dataset provides 108 entries, each containing metadata, general information, audience votes, transcript and summary.

3 System Design and Implementation

This is the second paragraph. Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

[**WEB:GNU:GPL:2010**]

3.1 Design

And after the second paragraph follows the third paragraph. Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

- Item 1
- Item 2
- Item 3
- Item 4

After this fourth paragraph, we start a new paragraph sequence. Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at

all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

3.2 Implementation

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

Algorithm 1 Test

```
1: procedure GET(a, b)                          ▷ The g.c.d. of a and b
2:     r ← a mod b
3:     while r ≠ 0 do                           ▷ We have the answer if r is 0
4:         a ← b
5:         b ← r
6:         r ← a mod b
7:     return b                                 ▷ The gcd is b
```

This is the second paragraph. Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

4 Evaluation

And after the second paragraph follows the third paragraph. Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

After this fourth paragraph, we start a new paragraph sequence. Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

4.1 Experimental setup

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

This is the second paragraph. Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the

look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

4.2 System evaluation

In this section we present the performance evaluation of the proposed system.

4.2.1 Performance and cost evaluation

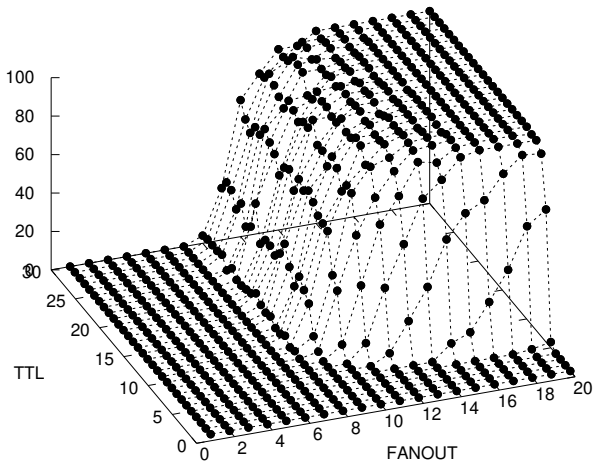

Fig. 4.1: This figure shows the percentage of complete disseminations.

And after the second paragraph follows the third paragraph. Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

After this fourth paragraph, we start a new paragraph sequence. Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet

First column	Second column	Third column
1	2	3
4	5	6
7	8	9

Tab. 4.1: Test table

and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

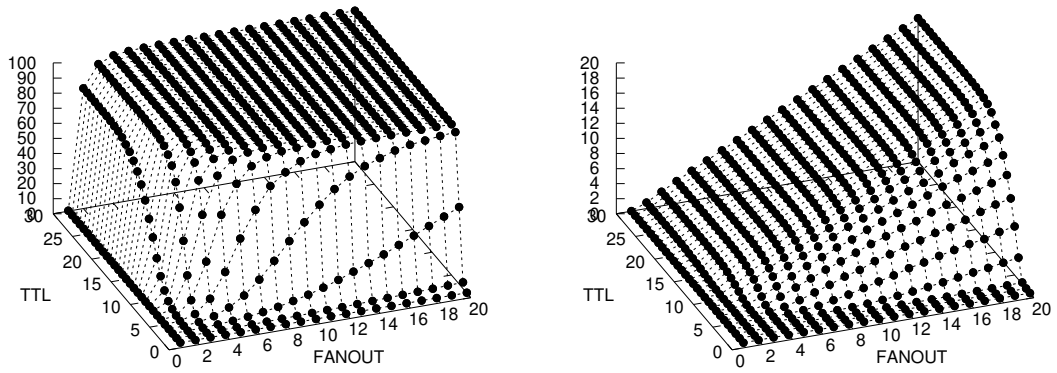

Fig. 4.2: These are the remaining two figures.

Note the detail in Figures 4.1 and 4.2.

Table 4.1 is a test table, listing all numbers from 1 to 9.

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

4.2.2 Qualitative evaluation

This is the second paragraph. Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the

original language. There is no need for special content, but the length of words should match the language.

And after the second paragraph follows the third paragraph. Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

4.2.3 Effect of other parameters

After this fourth paragraph, we start a new paragraph sequence. Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

5 Conclusions

This is the second paragraph. Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

As we see in Equation 5.1, this theory rocks… $\int_0^\infty \frac{x^4}{1+\frac{3}{x}} = 2 + i$

$$\int_0^\infty \frac{x^4}{1+\frac{3}{x}} = 2 + i$$

$$\int_0^\infty \frac{x^4}{1+\frac{3}{x}} = 2 + i \tag{5.1}$$

And after the second paragraph follows the third paragraph. Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

After this fourth paragraph, we start a new paragraph sequence. Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

List of Acronyms

API Application Programming Interface

WWW World Wide Web

List of Figures

List of Tables

List of Algorithms