

CSC/OR-591 Computational Social Choice Project

Evaluating the Performance of Voting Rules: A Comparative Study Using LLM-Generated Ballots

Ishan Hitesh Mistry

April 6, 2025

Problem Statement

Inspired by the paper *Generative Social Choice*, this project explores the potential of integrating large language models (LLMs) to address limitations in traditional social choice mechanisms. Current methods for capturing preferences often rely on voters ranking or selecting from a predefined list of candidates, questions, or items, resulting in significant constraints on expressibility and representation.

Key Challenges

- **Restricted Expressibility:** Traditional systems limit voters to choices from a fixed set of options. This rigidity may prevent voters from articulating preferences that lie outside or between the predefined choices, leaving them feeling underrepresented or dissatisfied.
- **Limited Scope of Representation:** The static nature of candidate lists restricts the range of preferences that can be captured, particularly when voters have nuanced opinions that do not align neatly with the provided options.
- **Granularity and Democratic Fairness:** While offering more granular options could make the process more democratic, it becomes practically difficult to balance comprehensiveness with usability, especially when the set of possible preferences grows large.

Mathematical Formulation

Let $P = \{p_1, p_2, \dots, p_n\}$ be a set of posts, where each post p_i consists of a tuple (t_i, d_i) , where:

- t_i is the title of the post, and
- d_i is the detailed description of the post.

Let $V = \{v_1, v_2, \dots, v_k\}$ be the set of possible verdicts, where each $v_i \in V$ corresponds to one of the predefined judgments such as "YTA", "NTA", "ESH", etc.

We aim to model a task where, given a post $p_i = (t_i, d_i)$, the goal is to predict the verdict $v_i \in V$ using a Large Language Model (LLM). The LLM \mathcal{M} takes the textual content of the post and generates the corresponding verdict. This task can be described as a sequence generation problem, where:

$$v_i = \mathcal{M}(t_i, d_i) \quad \text{where} \quad v_i \in V$$

In this context, the LLM is used to generate a verdict v_i based on the input post, without needing explicit feature engineering or label-based training typical of classification models. The model generates a response v_i from the set of verdicts V , conditioned on the input text (t_i, d_i) .

$$v_i = \mathcal{M}(t_i, d_i) \quad \text{where} \quad t_i \in T, \quad d_i \in D, \quad v_i \in V$$

Here, T represents the space of possible titles, and D represents the space of possible descriptions in the posts. The model is expected to predict the most likely verdict v_i based on the context provided by t_i and d_i .

LLM-Driven Solution

This project investigates how LLMs can expand the space of choices in preference aggregation processes. By generating dynamic and contextually relevant options, LLMs can provide voters with a richer, more expressive platform for articulating their preferences. Instead of being confined to static lists, voters could suggest or discover new options that better align with their views, enabling a more inclusive and nuanced representation of collective preferences.

Broader Implications

Integrating LLMs into social choice processes has the potential to transform how preferences are captured and aggregated. By addressing the limitations of fixed candidate lists and enhancing voter expressibility, this approach can foster decision-making systems that are more adaptive, democratic, and representative of diverse opinions.

Differences from the Paper

The approach I employed involves modeling human decision-making by representing individuals as agents based on their comments, and then simulating their ranking of a slate of statements. This method attempts to bridge the gap between machine-generated outputs and human decision-making by considering how an actual person might prioritize these statements. In contrast, the study I reviewed utilized actual human participants to validate the process, directly observing how people would interact with the voting system. Additionally, while my focus included analyzing the implications of various voting rules, the study primarily concentrated on ensuring fair representation in the slate of statements, without delving deeply into the consequences of different voting mechanisms.

Original CSC Concept

This paper focuses on the concepts of Justified Representation (JR) and Extended Justified Representation (EJR) in approval-based voting systems, ensuring that the selected slate of statements

reflects participants’ preferences in a proportionally representative manner. The system uses Approval Voting and Greedy Approval Voting to select a slate of k statements that most accurately represent the largest number of participants.

A key innovation in this framework is the use of generative queries, which leverage Large Language Models (LLMs) to generate new statements that are likely to receive the most approval from participants, rather than relying solely on existing ones.

The inspiration for this project comes from the concepts of Justified Representation (JR) and Extended Justified Representation (EJR) in approval-based voting systems, as discussed in the paper. These concepts ensure that a selected slate of statements accurately represents the preferences of participants. In my project, I leverage these principles by modeling and predicting how users would rank a set of statements based on their expressed opinions.

While the paper relies on actual human participants to validate the representation of preferences, my project takes this concept further by using Large Language Models (LLMs) to simulate human decision-making. By providing the LLM with both individual comment and post context, I can model how a participant might rank statements, ensuring the results align with the goals of Justified Representation.

Additional CSC Concept

The paper *Generative Social Choice* utilizes k -approval as a voting rule. However, in this project, I have taken an alternative approach to bridge the framework from the paper with additional methods, making it more accessible and adaptable.

For all voting rules, I employ a unified framework where preferences are represented in a preference matrix, and voting rules are applied consistently to this structure. This ensures that the methodology remains robust and extensible across different contexts.

In this project, I have extended this framework by incorporating LLM agents. These agents are tasked with generating a slate of statements dynamically. Once the slate is created, the preferences of all agents over these newly generated statements are obtained. This approach allows for a more flexible and expressive system, where preferences can be captured and aggregated beyond the limitations of predefined candidate lists.

Description of Computational Experiments

Dataset Extraction and Collection Process

What is the subreddit r/AITA? r/AITA, or *Am I the Asshole?*, is a subreddit where users share personal dilemmas, seeking judgments from the community about whether their actions were morally justifiable. Each post typically includes a title summarizing the situation and a detailed description, allowing the community to weigh in with one of six predefined verdicts. These verdicts include: **YTA** (You’re the Asshole), **YWBTA** (You Would Be the Asshole), **NTA** (Not the Asshole, and the other person is), **YWNBTA** (You Would Not Be the Asshole, and the other person would), **ESH** (Everyone Sucks Here), **NAH** (No Assholes Here), and **INFO** (Not Enough Information).

Why Was It Chosen? This subreddit was chosen because of its clear structure and inherent benefits for data collection. Posts are well-organized, and there is a mandate to include a final verdict for each situation, making extracting and analyzing labeled data easier. Furthermore, the format reduces the need for extensive cleaning compared to other datasets that might lack consistency or completeness.

What Problems Were Encountered with r/AITA? In the early stages of the project, the r/AITA subreddit was used to extract information, but several challenges emerged. A significant issue was the nature of the posts, as many regular contributors seemed to be either genuinely nice individuals seeking validation or people who simply wanted to be reassured. This often resulted in the verdict “Not the Asshole” (NTA) being the most common outcome, making it difficult to achieve a balanced distribution of verdicts. To address this, content had to be downloaded and filtered to ensure a diverse set of verdicts, but the process was hindered by rate limiting from Reddit’s API, which restricts the number of consecutive requests. Even when posts with varied verdicts were found, additional issues arose, such as insufficient participant engagement or low popularity, which could introduce bias into the dataset. To mitigate these challenges, strict rules were implemented in the data collection scripts, ensuring that only posts with fair representation and adequate response rates were included in the analysis.

How Were the Problems with r/AITA Solved? The solution to the challenges faced with r/AITA was to switch to a different subreddit, r/AITAFiltered. This subreddit is a curated list of “good” threads crossposted from r/AmItheAsshole, specifically selected to address the issues seen in the original subreddit. A bot automatically crossposts posts to r/AITAFiltered when they meet certain criteria: they are at least 18 hours old, ensuring that posts have had enough time for final judgments and flair assignments; they have at least 50 top-level judgment comments, ensuring a significant amount of engagement; and the majority judgment is not overwhelmingly dominant, with the votes being split at least 70/30, and ideally as balanced as 50/50. This ensures that r/AITAFiltered provides a more balanced distribution of verdicts, with a wider range of opinions. For simplicity, verdicts such as “Everyone Sucks Here” (ESH) and “You’re the Asshole” (YTA) are grouped together, as are “No Assholes Here” (NAH) and “Not the Asshole” (NTA). This curation effectively solves the problems faced in r/AITA, offering a fairer and more diverse dataset for analysis.

Attributes of the Dataset

The dataset extracted from r/AITA contains four main attributes: the **title**, which provides a concise summary of the situation; the **post ID**, a unique identifier for each post; the **description**, a detailed explanation of the dilemma; and the **final verdict**, which represents the community’s judgment.

EXAMPLE : DATASET ATTRIBUTES

Type: Post

Post_ID: 1gr10pe

Title: AITA for telling my friend she can get back in line...

Author: Anteaterreatingant

Timestamp: 2024-11-14 09:22:30

Text: So, long story short: I've had it with one of my friends. She has been rude and dismissive, and I finally told her she could "get back in line" if she wants to continue our friendship...

Score: 5597

Total_Comments: 1204

Post_URL: <https://www.reddit.com/r/AmItheAsshole/comments/1gr10pe>

Filtered_Verdict: The final verdict is: **No A-holes Here.**

Verdict: NaN

Popular: NAH

Data Cleaning Process

To ensure the dataset was suitable for further processing, several cleaning steps were undertaken. First, the text was cleaned to remove special characters such as `\n` and `\t`, which could interfere with natural language processing tasks. Links to other posts, gifs, and videos were also removed, as they did not contribute to the context or content of the dilemma. Additionally, extra spaces within the text were consolidated to create a contiguous block of text. This step was especially important to minimize the number of tokens processed by the language model, improving computational efficiency.

An essential part of the cleaning process involved removing rows attributed to bots, as they do not represent genuine human input. Notable bots like **Judgement_Bot_AITA** (49 occurrences), **AutoModerator** (35 occurrences), and **AITAMod** were identified and excluded from the dataset, resulting in the removal of 84 rows in total. These cleaning steps ensured that the data was both accurate and reflective of authentic community interactions.

Dataset Description

EXAMPLE : POST CONTENT

Title : 'AITA for yelling at my neighbour on the street at 4.30am?'

Text : 'My neighbour keeps idling his big, noisy truck at 4-5am outside my house, waking me and others up regularly. It is always for about 30 minutes at a time and most of the time he isn't even in the vehicle (he turns it on and goes back inside his house). Until now I have said nothing, because my spouse didn't want to rock the boat. This morning we both went outside and once my partner saw who it was immediately went back in. I couldn't contain myself and began yelling at him from outside my house - I angrily said \"stop doing this!\" possibly including a few expletives. AITA? I'm not proud of it but I am so tired. Between him and another repeat offender (a different issue which I have been to the authorities about), I am completely sleep-deprived. And it's so unnecessary on this neighbor's part - the truck is new and it's a balmy 80 degrees so he doesn't need to warm the engine... Note
- I have called the police and council before and as you might guess, NADA.'

Architecture of the experiment

The process begins by generating a slate of statements that encapsulate the diverse opinions expressed in the comments associated with a particular post. These summaries are crafted using a large language model (LLM) agent, which processes the text from the comments to produce concise and representative statements that capture the key viewpoints. This step ensures that the opinions are distilled into clear, actionable items that can be analyzed further.

Next, each commenter is treated as an individual agent, reflecting their unique perspective. This is done by modeling the commenter's position as an agent based on the content of their comment. Using this approach, we can ask the LLM agent to assess the commenter's preferences and determine a linear ranking of the slate of statements. Essentially, each commenter provides a preference order of the summarized statements, helping us understand how they value the different viewpoints presented.

With the ordered preferences of all the comment authors, we now have the necessary data to apply various voting rules. These rules can then be used to aggregate the preferences of all participants, determining the most popular or favored statements based on the collective judgment. This methodology leverages the power of LLMs for both summarization and preference ranking, enabling a nuanced and data-driven approach to analyzing public opinion.

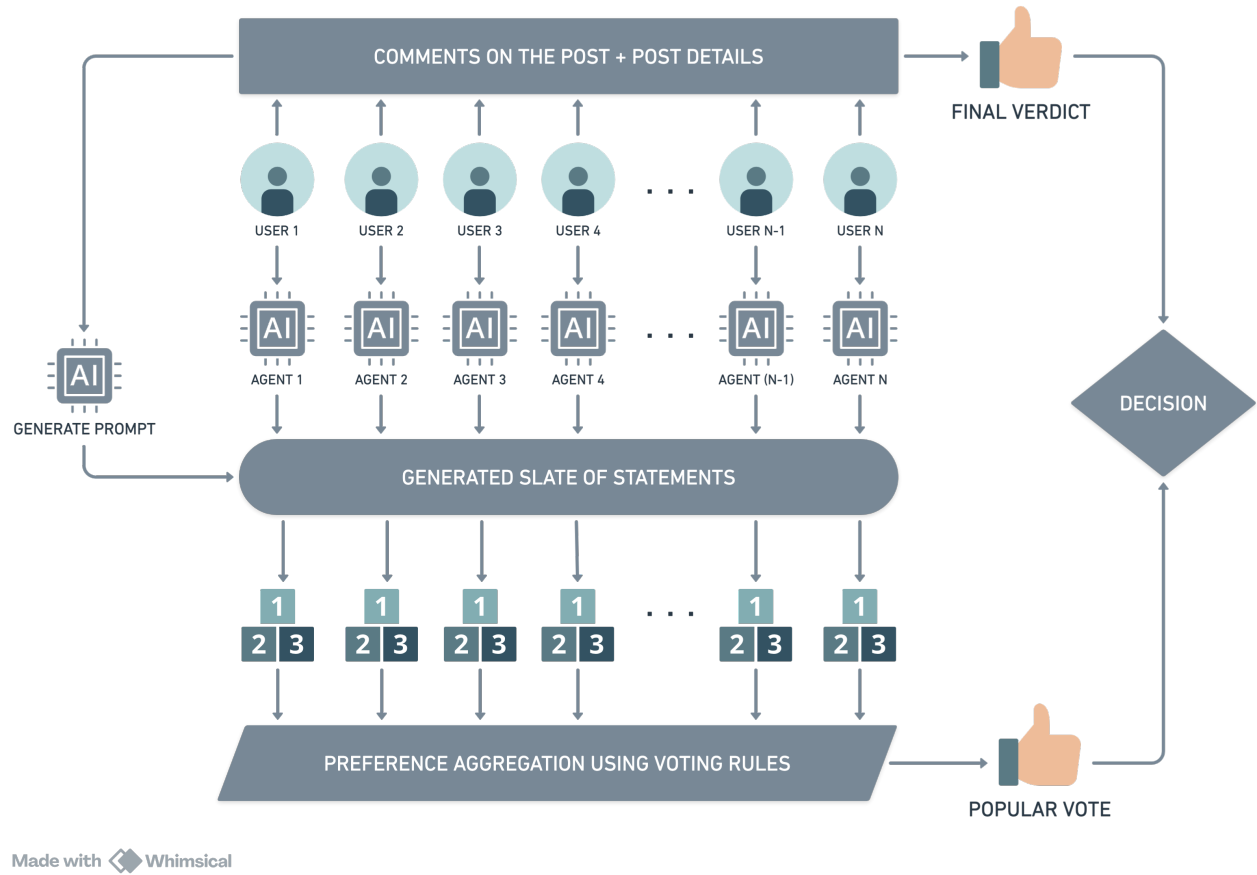


Figure 1: Enter Caption

Below is the continued example of how the LLM would generate a slate of statements based on the prompt. This prompt is used for generating a slate of statements by aggregating and summarizing all the comments for a particular post:

EXAMPLE : PROMPT : GENERATE SLATE OF STATEMENTS

"""

TASK :

Summarize the comments as CONCISELY as possible|no need for formal language or nice phrasing. Ignore all the verdicts that are present in the comments. There should not be any verdicts in the slate of comments.

DATA : Comments = {comments}

INSTRUCTIONS :

1. Determine the OPTIMAL(not too many and not too few) number of UNIQUE and SUCCINCT SUMMARIZED viewpoints you can find and give that many statements.
2. OUTPUT should in a JSON format, where the keys are strings of numbers and values are the summarized statements.
For example = dict('1' : 'Statement_1','2' : 'Statement_2')
3. ONLY return the output as a valid JSON(no preamble)
4. Usage of similar ideas and words in your summaries is a good indicator of duplicate ideas and such ideas should be avoided. Beware not every comment has a unique viewpoint.
5. DO NOT use your knowledge about the world; stick to what the participants said.Do not put quotes around your response.

"""

EXAMPLE : GENERATED SLATE OF STATEMENTS

```
{1: 'Boyfriend crossed a line by using a personal and sensitive topic to win a game.',
 2: 'He should have known better than to use a dead mom joke, even in a game meant to be offensive.',
 3: 'The game's goal is to be controversial, but that doesn't mean it's okay to hurt someone's feelings.',
 4: 'Boyfriend's apology was insincere, as he implied the OP was overreacting.',
 5: 'OP should set boundaries and communicate what is and isn't okay to joke about.',
 6: 'Boyfriend's actions show he's willing to hurt OP for personal gain, which is a red flag in the relationship.',
 7: 'OP should consider whether the boyfriend's behavior is a consistent pattern and if it's a sign of deeper issues.',
 8: 'The game is meant to be uncomfortable, but that doesn't mean OP should just accept being hurt.',
 9: 'Boyfriend's actions were not just about winning the game, but also about targeting OP specifically.',
 10: 'OP's feelings are valid, and the boyfriend should have been more understanding and empathetic.'}
```


Ballot Generation

PROMPT : GENERATE PREFERENCE ORDERS

```
AGENT_PROMPT_TEMPLATE = PromptTemplate.from_template(
    """
```

```
Given, the description and the comment as your stance, think like this person
and try to order ALL the statements from the slate in an order that goes
from MOST LIKED to LEAST LIKED. A random example would be as follows :, if the
example post description == "AITA for getting a rescue cat and giving it away
later on because it was not my cup of tea?".
```

```
Your example comment to this == "YTA, it was your duty to check and educate
yourself about the responsibilities that come with adopting a pet cat and how
sometimes they can be more difficult because they have been through abuse".
```

```
The example slate has 5 statements for example,
```

```
1 : You are not the asshole here because you were doing a good deed but did
not prepare for it.
```

```
2 : Well if it did not work out, no point in blaming people now.
```

```
3 : You are in the wrong here because you not only gave in to your impulses
but also were irresponsible with your duties.
```

```
4 : Cats look easy to maintain is not a good enough reason to get a stray
animal as your pet. You are not prepared to get one
```

```
5 : First, You need to interact and see if you are a cat person. Reach out
to people or friends who have pet cats.
```

```
-----
Given all this information you would have the following order of the
statements and the reasoning for the order.
```

```
ORDER = [3,1,4,5,2]
```

```
REASONING = Based on the comment you disagree and the poster is in the wrong.
Logically, you agree with 1,3 but then also agree with 4,5 a little because
that would mean being responsible. You do not agree with 2 so it goes last.
```

```
Now that the example is done. Here are the inputs for the task that you have
to perform. You are a commentor on a post. The post is from r/AITA with the
following description:
```

```
----
```

```
DESCRIPTION = {description}
```

```
----
```

```
You have put the following comment on the post as a reflection of what
you think :
```

```
----
```

```
COMMENT = {comment}
```

```
----
```

These are the summarized statements/opinions across all the other comments(including yours) that you have to use for the ordering and give reasoning for that order.

```
-----  
SLATE = {slate}  
-----
```

RULES TO STRICTLY FOLLOW :

1. DO NOT SKIP any statements. You NEED TO RANK all of the statements.
2. ONLY respond with ONE COMPLETE LIST OF ORDERING and the REASONING (based on your comment and slate) justifying that ordering.
3. FOLLOW the format from the example. You might have a different number of statements (10 or 15) so ADAPT accordingly.
4. Each RANK should be in the output EXACTLY ONCE.
5. Given above is JUST AN EXAMPLE for you. Do not use the same context, but try to understand how the task is to be done from this example.
6. DO NOT USE THE SAME SLATE or THE SAME ORDERING from the EXAMPLE, use the slate statements, description, and comment given to you after the example for the task.
7. DO NOT RETURN ANY EXTRA ITEMS except for the VALID ORDER and REASONING.

```
""")
```

Now that we have a slate of statements summarizing the opinions across the comments, the next step is to model each commenter as an agent. The goal here is to determine how the user would rank or order these statements if they were making a decision as an actual person, rather than simply expressing their opinion in a comment.

LLM Agent's Response

Below is a snippet of a random ballot generation process. In this process, the LLM agent is provided with a slate of statements derived from the opinions expressed in the comments across a post. Based on this input, the LLM generates two distinct outputs. First, it produces the ballot, which is essentially the linear order of the statements, reflecting how a commenter would rank them based on their opinion. Second, the LLM provides reasoning to justify the order it has selected. This reasoning helps clarify why the agent has chosen a particular ranking, taking into account the context provided by the post information and the individual commenter's perspective. The ballot and reasoning together simulate how an actual person might organize these opinions, allowing us to better understand and model decision-making in the context of the post.

OUTPUT : GENERATE PREFERENCE ORDERS

Processed row 448 with tokens = 230:
ORDER = [3, 9, 5, 1, 2, 4, 6, 8, 7, 10]

REASONING: Based on my comment, I am skeptical of the original poster's (OP) actions and believe they have a responsibility to inform their grandmother of their decision to move out. I agree with statements 3 and 9, which both criticize the OP for their lack of consideration towards their grandmother. I also agree with statement 5, which calls the OP ungrateful for not appreciating their grandmother's generosity.

I agree with statements 1 and 2, which suggest that the OP should give their grandmother notice before moving out, but I rank them lower because they do not address the OP's lack of gratitude.

I rank statement 4 lower because it suggests treating the grandmother like any other landlord, which does not take into account the personal relationship between the two.

I rank statement 6 higher than 1 and 2 because it emphasizes the urgency of the situation and the need for the OP to take responsibility for their actions.

I rank statement 8 lower because it suggests a specific time frame for giving notice, which may not be necessary in this situation.

I rank statement 7 lower because it suggests that the OP is not obligated to share personal details with their grandmother, which is not the main issue at hand. I rank statement 10 last because it suggests being honest with the grandmother, but does not address the lack of consideration and gratitude shown by the OP.

Progress saved to clean_data_pref_checkpoint.csv.

Results

The performance of the voting rules in this study is generally satisfactory. While the outcomes align reasonably well with expectations, there is still room for improvement to ensure even greater accuracy and robustness in the results.

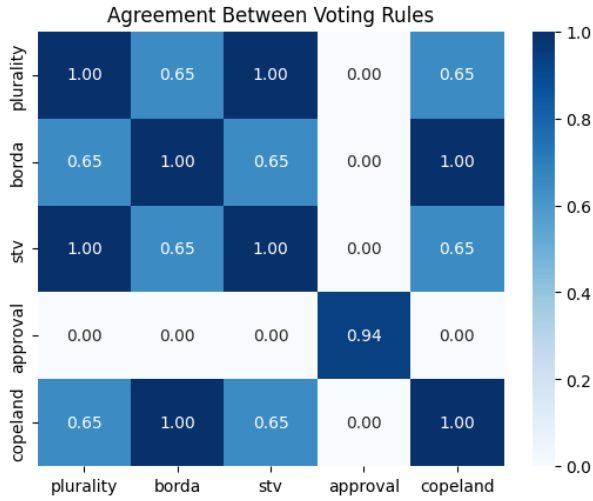
There are numerous avenues for improvement that could be explored in subsequent research. For instance, adjusting the weight given to different comments or experimenting with more sophisticated models of voter behavior could yield more nuanced results. Additionally, further exploration into alternative voting rules and their interactions with various post characteristics may help uncover new insights.

One particularly interesting observation from the study is the identification of a Condorcet winner in one of the posts. A Condorcet winner is a candidate or option that would beat every other alternative in head-to-head comparisons. In this case, all of the voting rules selected this particular winner, indicating a clear consensus among the different methods and suggesting that some voting rules may be more aligned with certain types of decisions or posts.

It is important to note that the "winner" column in the data represents the true winner, as determined by the popularity of the comment. This is based on the highest number of upvotes, providing an objective measure of which comment garnered the most support from the broader audience. This serves as a benchmark for comparing how well the different voting rules perform in terms of capturing the collective opinion.

Post ID	Popular	Plurality	Borda	Condorcet	STV	Approval	Copeland
1gi2awo	9	9	9	-1	9	NaN	9
1gi378c	3	3	1	-1	3	[8]	1
1gi5k6b	4	14	5	-1	14	[8]	5
1gi6rmt	3	10	7	-1	10	[1]	7
1giwkq5	1	1	1	-1	1	[1]	1
1giy3mj	7	7	6	-1	7	[2]	6
1gj6zkq	3	3	3	-1	3	[3]	3
1gjc9pm	3	3	3	1	3	[3]	3
1gjquini	3	3	3	-1	3	[8]	3
1gjvwc2	3	3	3	-1	3	[8]	3
1gjy9l8	3	3	6	-1	3	[4]	6
1gjyyhu	1	3	3	-1	3	[6]	3
1gk6tzm	1	3	3	-1	3	[6]	3
1gkbmkz	6	10	6	-1	10	[1]	6
1gkhfq7	1	6	6	-1	6	[1]	6
1gkhp7b	6	6	6	-1	6	[5]	6
1gkukkp	2	2	2	-1	2	[6]	2

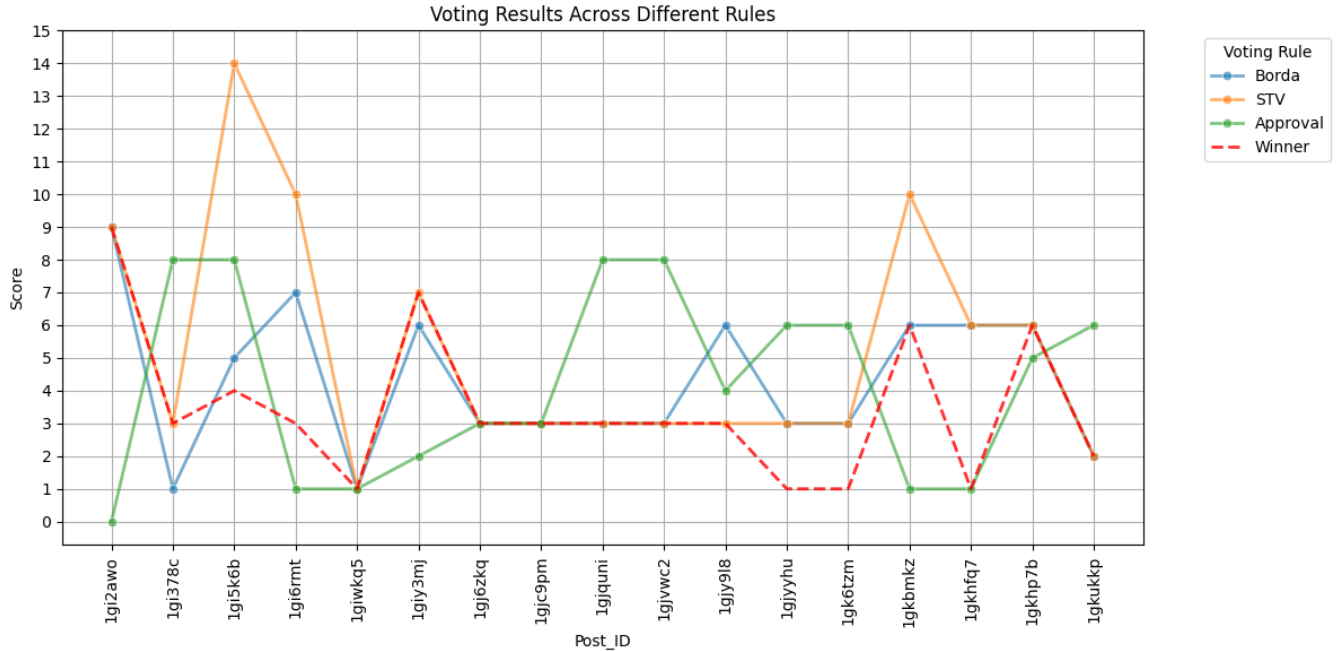
Table 1: Voting Results for Different Rules



Voting Rule	Accuracy
Plurality	0.6471
Borda	0.5294
STV	0.6471
Approval	0.2353
Copeland	0.5294

Figure 3: Accuracy for each voting rule

Figure 2: Heatmap of Voting Rules



Limitations

While this study presents promising insights, there are several limitations to consider:

- **Insufficient Data:** The analysis is constrained by the availability and size of the data. A larger dataset with more diverse comments would likely provide a more comprehensive evaluation of the voting rules and their implications.
- **LLM Context Window Issues:** One of the significant challenges faced during the project was the limitation imposed by the LLM's context window. Due to the size of the input data,

there were instances where not all the necessary content could fit within a single batch, leading to partial information being processed by the model. This constraint may have impacted the accuracy and quality of the generated statements and their subsequent analysis.

- **Computational Constraints:** The complexity of running LLMs and processing large datasets may result in high resource consumption and longer processing times, potentially limiting the scalability of the approach.
- **Simplification of Voting Rules:** The study focuses on basic voting rules, which may not fully capture the intricacies of more complex real-world conversations.

Conclusion

This study provides valuable insights into the performance of various voting rules in capturing the collective opinions of participants. While the results are generally satisfactory, there is significant potential for improvement in the accuracy and applicability of these voting systems. The identification of a Condorcet winner across all voting rules suggests a strong consensus in certain cases, while the varying accuracies across rules highlight the need for further exploration. Future research could focus on refining the models, exploring alternative voting mechanisms, and examining the role of post characteristics in shaping voting outcomes.