AI-assisted Policy Making using Generative Social Choice

Suhas Vundavilli
BTech 3rd Year
Indian Institute of Science
suhasv@iisc.ac.in

Abstract

This report presents an **AI-assisted framework for participatory policy drafting** by combining natural language processing and computational social choice. Drawing inspiration from the paper **"Generative Social Choice"** by [1], the methodology integrates free-text public input with generative models and voting rules to construct a representative set of policy statements. Public opinions are collected through open-ended surveys, and major thematic clusters are identified using hierarchical clustering based on textual embeddings. A large language model is then used to generate candidate policy statements that reflect these themes. To ensure factual accuracy and democratic representativeness, the **Balanced Justified Representation** (BJR) voting rule is implemented to select a subset of statements approved by diverse segments of the participant pool. Approvals are inferred using cosine similarity between participant responses and candidate statements, either via TF-IDF or semantic embeddings. The selected statements are further validated through participant feedback. This work demonstrates a scalable and transparent mechanism for crowdsourced policy generation, combining machine learning with formal fairness guarantees.

1 Introduction

Social choice theory is the field of mathematics, economics, and political science that studies the aggregation of individual preferences towards collective decisions. The typical social choice setting involves a *small, predetermined set of alternatives* (such as candidates in an election) and a set of participants who specify their preferences regarding these alternatives, often in the form of rankings. When the set of alternatives is small and well-defined, social choice theory provides a rich set of tools for analyzing the properties of different voting rules and their implications for collective decision-making. However, in many real-world scenarios, the set of alternatives is *large, open-ended, and nuanced*, making it difficult to apply traditional social choice theory.

As an example, we can look at Citizens' Assemblies in Western democracies [2], which are groups of citizens selected to deliberate on a specific issue and make recommendations to policymakers. These assemblies often face challenges in defining the set of alternatives for deliberation, as the issues at hand are complex and multifaceted. Another challenge is the need to ensure that the presented alternatives are inclusive and representative of the diverse perspectives within the assembly, and those affected by the decision.

In response to the rapid progress of generative AI tools, and Large Language Models (LLMs) in particular, private corporations such as Meta [3] and OpenAI [4] have begun experimentation on aligning AI values with democratic processes, by asking open-ended questions such as *"how far [...] personalization of AI assistants like ChatGPT to align with a user's tastes and preferences should go?"*.

35th Conference on Neural Information Processing Systems (NeurIPS 2021).

Such developments have led to the emergence of a new field known as **Generative Social Choice**, which aims to infuse the flexibility and immense generative capability of LLMs into the rigorous framework of traditional social choice theory.
In the view of the authors of [1], the challenges of using traditional social choice theory to answer open-ended questions essentially boils down to two fundamental obstacles, both of which have been shown to be surmountable with the help of LLMs.

- *Unforseen Alternatives:* When we look at traditional social choice theory, the set of choosable alternatives is explicitly defined and static. One such example is the Brexit referendum of 2016, where there were only 2 clear choices: maintain the status quo, or split from the EU. Due to the lack of an intermediate options in between, voters were forced to pick on of these options, even if they might not have fully captured their preferences. By contrast, LLMs have the power to generate a virtually infinite number of alternatives, some of which may not be obvious to humans but help find a common middle ground. In principle, the possible outcomes of an LLM-augmented democratic process may span the universe of all relevant outcomes for the problem at hand.

- *Extrapolating Preferences:* In classical social choice thoery, the agents specify their preferences in a rigid format. Normally, the agents would evaluate each alternative individually on its own merit, or use a voting rule to rank the alternatives[1]. This approach is, thus, inherently flawed when alterntives which were not previously anticipated are introduced, and hence, not elicited. In such scenarios, LLMs can be of great help by allowing all participants to implicitly specify their preferences by articulating their ideas and opinions in natural language. The LLM can then act as a proxy for the participant and predict their preference response to any alternative, whether previously seen or not.

1.1 A framework for Generative Social Choice

It has been made clear until now that LLMs can be used to help overcome the challenges of unforeseen alternatives and preference extrapolation. However, the pairing of social choice theory and LLMs might not be as straightforward as this report makes it seem, since social choice is based on a foundation of rigorous guarantees, while LLMs are notorious for their imperviousness to theoretical analysis.
So, to address this difficulty, the authors of [1] propose a framework for Generative Social Choice which breaks the desing of democratic processes into the following two interacting components:

- *First Component:* **Guarantees with perfect queries**. Assume that the LLM is an oracle that can precisely answer certain types of queries, which may involve generating new alternatives in an optimal way or predicting agents' preferences. Once appropriate queries have been identified, the task is to design algorithms that, when given access to an oracle for these queries, provide social choice guarantees.

- *Second Component:* **Empirical validation of queries**. Assuming the LLM to be a perfect oracle is helpful for guiding the design of a democratic process, but of course not an accurate reflection of reality. In the second component, the task is to implement the proposed queries using calls to LLMs, and to empirically validate how well these implementations match the queries.

Naturally, the two components are interdependent. The theory of the first component identifies queries that are useful for social choice and must be validated empirically. Meanwhile, the experiments show which queries can be answered accurately in reality, which raises the question of the guarantees the algorithms relying on these queries might provide.
The biggest benefit of using such a framework is that the theoretical results we derive from it are future-proof, in the sense that as LLMs contiue to rapidly improve, their reliability in answering queries can only increase, providing even more power to our LLM-based aggregation methods.

[1] This is the case, for example, in multi-winner elections.

2 Literature Review

2.1 Oppurtunities and Risks of LLMs for scalable deliberation with Polis, 2023

This paper [2] discusses the oppurtunities and risks of using LLMs for scalable deliberation with Polis, a platform for online deliberation.
The oppurtunities identified include topic modelling, summarization, moderation, comment routing, identifying consensus and vote prediction, all of which are, to some degree, already being used in multiple online platforms.
The most relevant of these experiments to this report are the ones on *summarization* and *vote prediction*, which shall be used in the later sections of this report.
In the near future, our democratic processes as a whole might be able to serve in the summarization role as outlined by the authors of [2], for which they do not propose any specific algorithms or perform any systematic experiments.

2.2 Elicitation Inference Optimization for Multi-Principal-Agent Alignment, 2022

This paper [3] proposes a framework of integrating an LLM with a latent factor model in order to predict preferences.
In a broader sense, the authors of [3] believe the paradigm of *virtual democracy* facilitates automated decision on ethical dilemmas by learning the preferences of relevant stakeholders and predicting their preferences over current alternatives, and then aggregating these predicted preferences.
Some example papers which make use of classical machine learning approaches to apply this paradigm to various fields are [5], [6] and [7].
One common thread among the above papers is their aim to predict preferences from a fixed set of alternatives - no new alternatives were generated.

2.3 Fine-tuning language models to find agreement among humans with diverse preferences, 2022

In this paper [8], the authors fine-tune an LLM to generate a single consensus statement for a specific group of people, based on their written opinions and ratings of candidate statements.
Then, reward models are trained to capture the individual preferences, and a social welfare function is used to judge how acceptable a certain statement is to the group.
The main difference between the work done in this paper [8] and the work done in both [1] as well as this report, is that we do not attempt to find a single statement that is acceptable to all participants, but rather a set of diverse statements that are representative of the participants' preferences.
A more fundamental difference is that we view our experiments as an instance of a broader framework that allows for a systematic investigation of the types of queries an LLM can perform and the theoretical guarantees they provide.

2.4 Futher Reading

- **Justified representation in approval-based committee voting, 2017** - This paper [9] has provided the base for justified representation in approval-based committee elections, which is the voting rule we will be using in our experiments.

- **Representation with incomplete votes, 2023** - This paper [10] provides a study on representation axioms in a statement-selection context. The key technical challenge in this work is the access to only partial approval votes.

It must be noted that all previous papers in this literature assume a fixed set of alternatives in a non-generative setting, while the work done in this report is based on a generative setting, where the set of alternatives is not fixed and can be generated by an LLM.

3 Multi-Winner Voting

3.1 Balanced Justified Representation

Setup: We have a set of n agents, represented by the set $N = \{1, 2, \ldots, n\}$ and the universe of (well-formed and relevant) statements given by $\mathcal{U}$.
Each agent $i \in N$ has a unitility function $u_i : \mathcal{U} \to \mathbf{R}$, which assigns a real utility to each statement.
We define an *instance* of the statement-selection process to be a tuple consisting of $N, \mathcal{U}, \{u_i\}_{i \in N}$ and a slate size $k \in \mathbf{N}_{\geq 1}$.
Now, we shall define balanced justified representation (BJR) as follows:
Definition 1 (Balanced Justified Representation)
A slate W is said to satisfy **Balanced Justified Representation** (BJR) if there exists a function $\omega : N \to W$ matching agents to statements in a balanced way[2], such that there exists no coalition $S \subseteq N$, a statement $\alpha \in \mathcal{U}$ and a threshold $\theta \in \mathbf{R}$ such that

1. $|S| \geq n/k$,
2. $u_i(\alpha) \geq \theta$ for all $i \in S$, and
3. $u_i(\omega(i)) < \theta$ for all $i \in S$.

In other words, if there exists a coalition of agents that is

1. large enough to be representative of the population and deserve a statement on the slate,
2. has cohesive preferences[3], and

then, the coalition must not be ignored[4],

The concept of BJR given here is connected to various axioms in social choice theory. Relaxing the balanced condition of BJR, thereby simply matching agents to their most preferred statements, can yield some interesting relationships. In the context of approval utilities, this relaxation is in alignment with the **Justified Representation** (JR) axiom given in [9]. For our framework of general cardinalities, this new relaxed version is implied by both **Extended Justified Representation** (EJR) and **Full Justified Representation** (FJR), as defined in [11].

To show the need for a new balanced-matching-way way of justified representation is shown through the following two examples:

3.1.1 Example 1

	α	α'	β	β'
u_1	1	1	0	0
u_2	1	1	0	0
u_3	0	0	1	1

Table 1: Utility matrix with $k = n = 3$.

In this example, $k = 3$ statements must be chosen, and two-thirds of the agents (namely 1 and 2) approve of α and α', and the other one-third (namely 3) approve of β and β'.
As seen in Example 3 of [9], the slate $\{\alpha, \beta, \beta'\}$ satisfies the constraints of JRm and thus by extansion, those of BJR with relaxed unbalanced matchings.
However, we run into a problem: two-thirds of the population is represented by only one-third of the slate, and vice versa.
This occurs because JR cannot rule out this form of unproportionality as each member of the two-thirds bloc is already represented by some statement they approve, and JR does not allow agents and coalitions to formulate any claims to representation beyond that point.

[2]That is, each statement on the slate is matched to $\lfloor n/k \rfloor$ or $\lceil n/k \rceil$ agents.
[3]That is, there is a statement in $\mathcal{U}$ for which all agents in coaltion have utility atleast θ
[4]That is, atleast one agent in the coalition must be assigned a statement on the slate with utility at least θ

This is a clear violation of the BJR axiom, which requires that the representation of the population and the slate be balanced.
However, this problem can be mitigated by simply using either EJR or FJR, which assume that an agent may be represented by more than one statement on the slate.

3.1.2 Example 2

	α	α'	β	β'
u_1	3	0	2	2
u_2	0	3	2	2

Table 2: Utility matrix with $k = n = 2$.

In this example, we are required to choose two statements for two agents.
It can be observed here that each agent $i \in \{1, 2\}$ has a statement α_i which is very specific to i and has a high utility. Thus, a slate consisting of only these two statements would be a good pick since it represents the specificity of agents' preferences to the highest degree.
Indeed, the slate $\{\alpha, \alpha'\}$ is the only slate here that satisfies all the constraints of BJR, and thus must be picked.
EJR and FJR fail in this particular case since they prefer to represent both agents jointly by two less specific statements (namely β and β') rather than each agent individually by a specific statement.

Thus, the axiom of BJR is able to rule out disproportional representation in the first case, and is able to successfully enforce a higher degree of specificity in the second case, in comparision to the other representation axioms.
On closer inspection, we see connections between BJR and the notion of *fully proportional representation* given by [12]: "voters should be segmented into equal-sized coalitions, each of which is assigned a representative, such that the preferences of voters are as closely as possible reflected by the representatives of their segment."

4 Summary of Pilot on Chatbot Personalization

In this section, we shall summarize the pilot study on Chatbot Personalization as described in [1].

4.1 Pilot Description

100 participants were recruited through the online platform Prolific, with all of them being residents of the United States and diverse with respect to age, gender and race.
The participants were asked to fill out a survey on chatbot personalization.
To start off, the participants were shown background information and asked questions on whether the chatbot must personalize its response in each of three example scenarios.
Later, participants were instructed to articulate their thoughts on chatbot personalization, by answering four questions in natural language on topics such as trade-offs of personalization, guardrails they would like to see imposed on the personalization process and arguments in favour of as well as against their proposed rules.
Participants were further asked to rate six example statements, which had been generated using GPT-4 and without knowledge of participant responses. Ratings were given on a five-point scale.
Using these responses, a slate of 5 representative statements were extracted using the BJR voting rule described in previous sections.
A new set of 100 participants were now shown this representative slate and asked to rate the statements, in a similar fashion to the previous survey.

4.2 Generated Slate

The generated slate was a set of 5 statements, which are as follows:

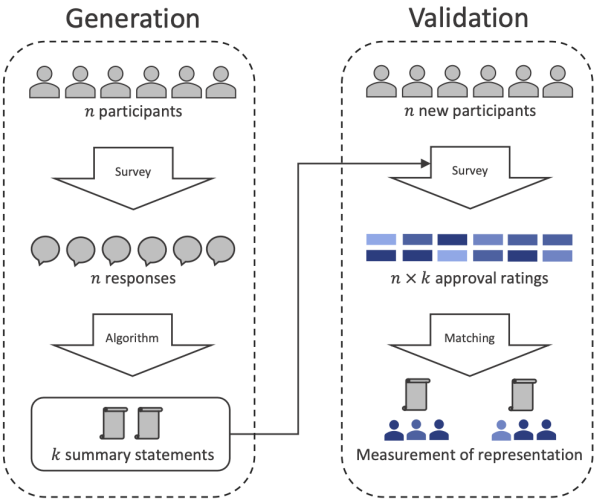

Figure 1: Overview of the pilot process, where $n = 100$ and $k = 5$. Taken from [1]

S1. The most important rule for chatbot personalization is to ***give users control over the extent of personalization and the data supplied***. This rule is crucial as it ensures user autonomy, ***privacy***, and a personalized experience. For instance, a user could choose to share their dietary preferences with a health chatbot for tailored advice, while opting not to disclose sensitive health data.

S2. The most important rule for chatbot personalization is to always ***give users the choice whether the AI chatbot can remember their data or not***. This rule is crucial because it ***respects the user's privacy*** and gives them control over their own data. For instance, a user might prefer a chatbot not to store any data about their past travels, thus avoiding unsolicited vacation suggestions.

S3. The most important rule for chatbot personalization is to always ***prioritize user privacy and data security***. This is crucial because it ensures the protection of sensitive user information, thereby building trust and promoting responsible AI use. For instance, a chatbot providing personalized health advice should only ***collect and use data with explicit user consent***, and should implement robust measures to prevent unauthorized access or data breaches.

S4. The most important rule for chatbot personalization is to ***avoid providing false or misleading information***. This rule is crucial because it ensures the reliability and trustworthiness of the chatbot, which is essential for user engagement and satisfaction. For instance, if a user asks a chatbot for medical advice, providing accurate information could potentially save lives.

S5. The most important rule for chatbot personalization is to ***emphasize privacy*** and require ***user consent for data collection***. This rule is crucial to ensure personal security and mental health protection. For instance, a health bot providing personalized services can offer tailored care, but without proper privacy measures, it risks violating user privacy.

By studying the above statements, we can see the emergence of three main themes intertwined with each other:

- **Privacy and data security:** Four out of five statements stress the importance of privacy and of preventing chatbot data from being used in other contexts.
- **User control:** Four out of five groups believe that it is essential that users have granular control over which of their data are stored and used for personalization.
- **Truthfulness:** The third statement's primary concern is that chatbots should never provide inaccurate or misleading information.

It is interesting to note that the same themes have been repeated in multiple statements. However, it is worth highlighting that different statements represent different nuances of the same idea. For

example, S5's concern on privacy and user control is in accordance with security and mental health concerns, which seems highly specific when compared to the more vague and generic description seen in S2. It must also be noted that S1 is of the opinion that privacy appears as only one of the many underlying values served by user control, and stresses on the importance of user control not just at the time of data collection, but also on the level of personalization when the chatbot is subsequently used. Finally, the remaining statement, S4, stresses that chatbot personalization should not go so far as to compromise the chatbot's truthfulness.

4.3 Representation of slate in generation sample

Given the novelty of the process and the central role LLMs play in this project, it is of paramount importance to verify that the slate of reoresentative statements are indeed faithful to the participants' opinions and not based on LLM hallucinations.
In order to do this, the authors of [1] manually inspected and hand-labelled the responses of the generation sample in order to see how the process arrived at the slate from the participants' responses. On doing so, it was foound that privacy and data security as well as user control were central themes in many of the free-text responses of the participants: 61 of the 100 participants touch on privacy and data security in their statements, 38 suggest user control, and 72 bring up at least one of these two topics.
The number of 72 participants who touched on privacy and data security and user control alone can plausibly justify that these themes take up 80% of the slate. Moreover, this number does not yet count agents who expressed agreement with these themes outside of the free-form responses.
In light of these observations, representing 80% of agents with a statement about privacy and data security and user control seems like a reasonable choice.

4.4 Representation of slate in validation sample

By the ideals of proportional representation, each statement in the slate should accurately represent one-fifth of the US population. To verify this, the authors of [1] matched the participations of the validation group to the statements in the slate which maximize the sum of participants' rating levels for their assignment.
By studying the given chart in the below figure, we can see that 75% of the participants say that their assigned statement "perfectly" captures their full opinion on chatbot personalization, and an additional 18% of participants say it "mostly" captures their full opinion. Only 7% of participants feel only "somewhat" represented or less.
Hence, the vast majority of participant opinions are represented accurately by our slate of statements. A remarkable observation to be made is that none of the agents had a higher rating for a statement different from their assigned statement, which means that the requirement to assign an equal number of agents to each statement is not a binding constraint.
Naturally, it is important to closely inspect the minority of 7 agents who feel relatively badly represented by their assigned statement, since their responses could potentially reveal viewpoints missing from our slate. Though the free-text explanations given with the ratings are typically short, they allow us to understand what the seven participants dislike about the selected statements. While certain themes occur repeatedly among these seven participants,24 their reasons for feeling relatively unrepresented are eclectic. Since proportionality axioms like BJR only guarantee representation to large, cohesive groups, these responses also give us no reason to doubt the representativeness of our slate.

4.5 Results and Discussion

This pilot study has successfully shown that a democratic process for selecting representative statements is not just a hypothetical exercise, but one of practical relevance.
In terms of reliability, GPT-4 does sometimes produce imprecise and/or unpopular opinions. The current implementation of the generative query increases robustness by generating multiple candidate statements with different approaches, and selecting the best from them. However, our process has yet to be hardened against malicious participant input, such as prompt injections meant to sway the generative queries in particular directions [13].
Another major issue in need of mitigation is the well-known biases that LLMs possess against specific groups of people and perspectives. This could cause impediments to the goal of inclusive

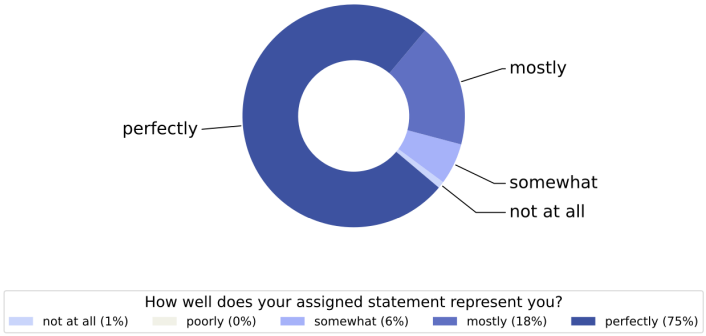

Figure 2: Ratings of participants from the validation survey for their assigned statement. Taken from [1]

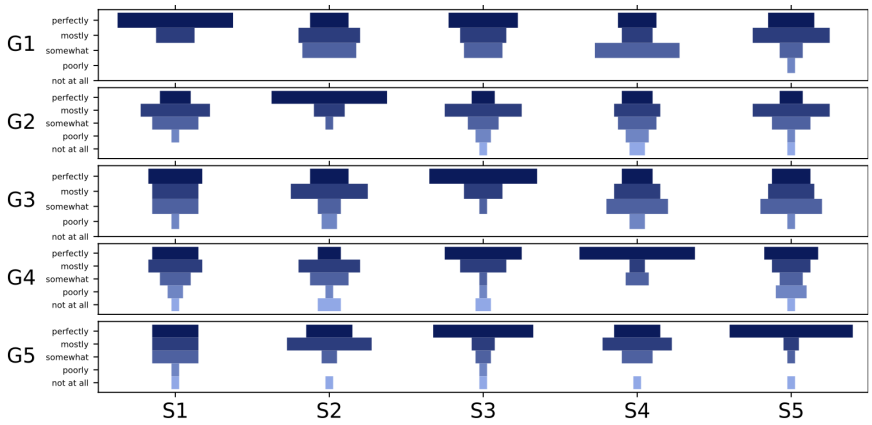

Figure 3: Agreement of participants in different groups with each of the statements. Each row corresponds to a group; for example, G1 represents the 20 participants assigned to statement S1. For each group, we plot the frequencies of rating levels given by members of this group to statements S1 through S5. Taken from [1]

and representative decision-making.
The biggest challenge the authors feel is the lack of transparency inherent to any process involving LLMs. This issue, and the resulting threat to the legitimacy of the process, could be alleviated by adding participatory stages. A simple example of this is the pilot's validation round, in which a fresh sample of participants, without any involvement of the LLM, demonstrated a fit between the generated slate and public opinion.
Enhancing democratic processes with LLMs opens up new points on the legitimacy-scalability tradeoff curve, which could, in practice, enable new forms of collective decision-making.

5 Methodology Overview

5.1 Data Collection

A primary survey was conducted to collect opinions and thoughts on the topic of our country's educational policy. This was achieved through a Google Form sent to students, both in and outside of the Indian Institute of Science.
This survey mainly consisted of open-ended questions on various facets of the educational policy, and encouraged answers in natural language. This survey, alongwith the data collected, is a part of

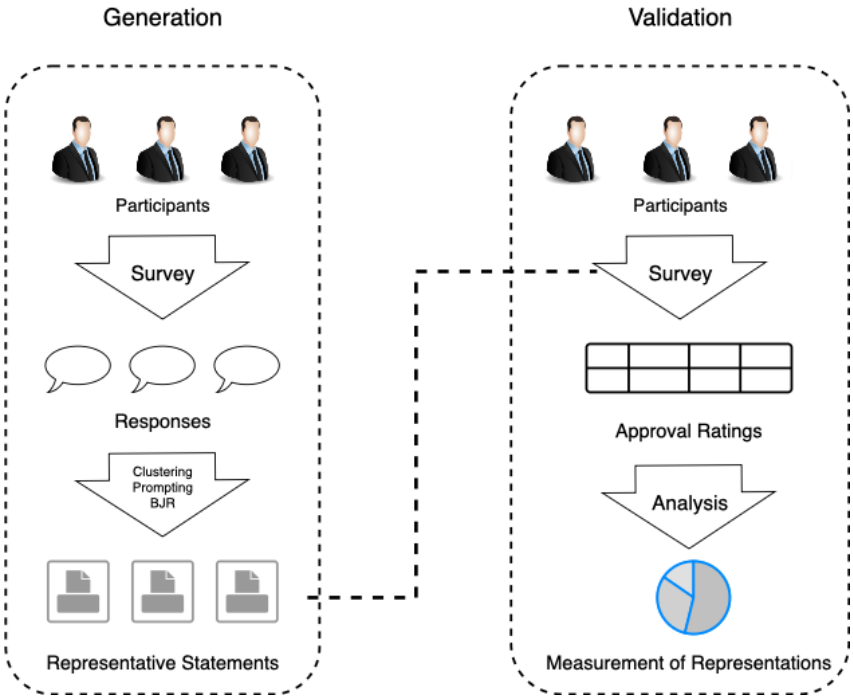

Figure 4: Overview of the methodology used in this report

the GitHub repository of this project, linked at the end of this report.

5.2 Theme extraction

The data colected from the survey was then used to extract major themes among the responses. The responses were first pre-processed to remove irrelevant answers, and the remaning were vectorized using `TfidVectorizer`, which is a part of the `sklearn` library.
Once the vectorization was complete, hierarchical clustering was performed on the responses using `AgglomerativeClustering` from the `sklearn` library, with the `cosine_distance` metric.
The reason for not choosing `KMeans` clustering is the very large size of responses in the dataset, which would have made the clustering process very slow and ineffective.
After much trial and error, the number of clusters was set to 8, which was found to be the most optimal number of clusters. In doing so, each cluster turned out to be roughly of the same size, alongwith a good mix of responses from different participants in each cluster.

5.3 Statement Generation

Representative statements for each cluster were generated by multi-shot prompting the LLM. The prompt was designed to ask the LLM to generate a statement that best represents the opinions of the participants in each cluster.
The prompt used is a part of the Appendix.
In all three LLMs were used to generate the statements: `GPT-4o`, `Gemini 2.0 Flash` and `Claude 3.7 Sonnet`.
Thus, the final set of candidate statements was of size 3×8, with 3 statements for each of the 8 clusters.

5.4 Statement Selection

BJR was used to select the final set of statements from the candidate statements generated in the previous step.
Since we do not explicitly know the participant ratings for the newly generated statements, this was achieved via generating approval ratings for each of the candidate statements.
In order to do this, the cosine similarity of each candidate statement was computed with each of the original responses for every agent, in this case a participant.
If the cosine similarity was greater than a certain threshold, the candidate statement was assigned a rating of 1, else 0.
Finally, the rating of each agent for each candidate statement was the sum of the ratings assigned to it.
Once the ratings were assigned, the BJR voting rule was applied to select the final set of statements.
In all, 12 statements were chosen from the set of candidate statements, 4 statements per LLM.

5.5 Statement Validation

Another survey was conducted to validate the statements generated in the previous step.
Participants were provided with a Google Form, with 4 sections. Each section contained 3 statements pertaining to a similar theme, one from each LLM. Participants were then asked to rate each statement on a scale from 1 to 5 and also rank the statements in each section.

6 Results and Discussion

The final representative statements for each LLM are given below:
GPT-4o:

- S1 India's education policy should focus on elevating public schools through better infrastructure, trained and fairly compensated teachers, inclusive decision-making, and updated curricula that emphasize critical thinking, real-world skills, and digital access—while fostering healthy collaboration and accountability between public and private institutions to ensure equitable and high-quality education for all.
- S2 India's education policy should focus on fostering strong collaboration between academia and industry to promote hands-on, project-based learning and develop adaptable, real-world skills such as communication, critical thinking, and digital literacy.
- S3 India's education policy should focus on increasing awareness, ensuring access to quality standardized resources like government-approved video lectures, promoting meaningful project-based learning, and establishing a transparent body to uplift or reform substandard institutions.
- S4 India's education policy should focus on revamping primary education to prioritize core subjects like mathematics, science, and economics, while eliminating outdated content and ensuring better teaching quality.

Gemini 2.0 Flash:

- S1 India's education policy should focus on improving the quality and accessibility of public education by investing in infrastructure, technology, teacher training and compensation, curriculum updates emphasizing critical thinking, problem-solving, and relevant skills, while also establishing better oversight and accountability for both public and private institutions through feedback mechanisms and potentially public-private partnerships to ensure equitable and quality education for all.
- S2 India's education policy should focus on enhanced collaboration with industry to integrate practical skills, provide real-world exposure through internships and project-based learning, and ensure the curriculum remains relevant and equips graduates with adaptable skills for the job market.
- S3 India's education policy should focus on strategically integrating AI and digital tools to personalize learning, enhance teaching effectiveness, and provide flexibility, while also

prioritizing teacher quality through better compensation, rigorous selection, and continuous training, and shifting towards more practical, competency-based assessments that emphasize completion over competition.

S4 India's education policy should focus on increasing awareness and guidance, potentially through accessible resources like cheaper internet, awareness campaigns, recorded standard lectures, government-approved video content, and establishing a quality board to address underperforming institutions and ensure standardized education.

Claude 3.7 Sonnet:

S1 India's education policy should focus on improving the quality of public schools through better teacher compensation, enhanced infrastructure, critical thinking-centered curricula, and digital literacy, while establishing stronger accountability mechanisms that incorporate feedback from teachers, students, and parents to create an accessible and forward-looking educational system that provides opportunities for all children regardless of socioeconomic background.

S2 India's education policy should focus on strengthening collaboration between academic institutions and industry through internship programs, project-based learning, and co-designed curricula that balance theoretical knowledge with practical skills development, ensuring students graduate with relevant competencies for the modern workforce.

S3 India's education policy should focus on increasing awareness through internet-accessible resources like standard-based video lectures, providing better guidance for students, developing meaningful project-based learning, and establishing a quality oversight board to improve underperforming institutions.

S4 India's education policy should focus on goal-oriented learning that aligns curriculum with socioeconomic needs, emphasizes practical outcomes for employment, and adapts regularly to changing job market requirements, all while fostering student purpose and self-care.

As seen in the above statements, it is clear that the statements generated by the three LLMs have major overlapping themes within them. For example, S1 of every LLM is concerned with the quality of public schools and improving the overall curriculum, while S2 is focussed on collaboration between academia and industry in order to prepare the students adequately to join the workforce. Similarly, S3 delves into the role of technology, and AI specifically, in the education system, while S4 is concerned with the need for a more practical and goal-oriented curriculum.
These statements are indeed representative of the clusters they were generated using, and thus can be used to represent the opinions of the participants in the survey.

The distribution of ratings given by the participants in the validation survey is shown in the below figure.

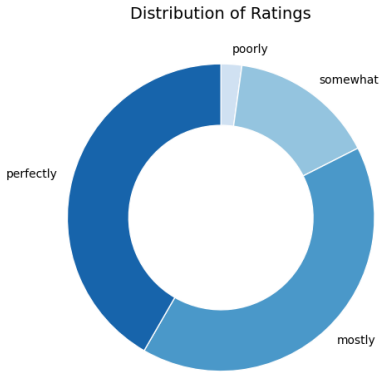

Figure 5: Ratings of participants from the validation survey for the given statements.

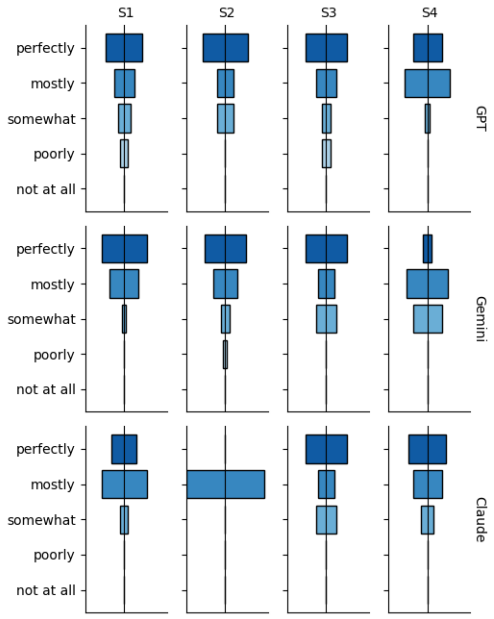

Figure 6: Agreement of participants for each statement presented.

As can be seen in the above plots, the ratings given by the participants are quite high, with most of them rating the statements as either "mostly" or "perfectly" capturing their opinion.
This is a clear indication that the statements generated by the LLMs are indeed representative of the participants' opinions, and thus can be used in the process of generating representative statements.
Another interesting fact is the general indifference of participants towards the statements generated by different LLMs. This may be interpreted as a sign of almost equal capability of all the LLMs used in this project.
When asked to rank similar statements from different LLMs, participants were indifferent to the LLMs and rated the statements based on their content rather than the LLM that generated them.
Gemini did have a slight edge over the other two LLMs in regards to S1, wheras Claude had a slight edge over the other two LLMs in regards to S3.
When it came to S2 and S4, the ratings were almost equal for all three LLMs.

7 Strengths and Limitations

7.1 Strengths

Both, the pilot study conducted in [1] as well as the study conducted in this report, have shown that the use of LLMs in democratic processes can be a powerful tool for generating representative statements.
The use of BJR as a voting rule has also shown to be effective in selecting the final set of statements from the candidate statements generated by the LLMs, which are themselves representative of the participants' opinions.
This clearly evident by the high ratings participants have given to the generated statements.
The use of LLMs in this process has also shown to be effective in generating diverse and nuanced statements, which are representative of the participants' opinions.
This process is thus democratic, efficient and scalable to a large number of participants, and can be used in various contexts to generate representative statements, and could be a gamechanger in the field of policymaking.

7.2 Limitations

The biggest limitation of this process is the lack of transparency in the LLMs, which can lead to biases and inaccuracies in the generated statements.
This is a well-known issue with LLMs, and it is important to be aware of this limitation when using them in democratic processes.
Another limitation is the potential for LLMs to hallucinate statements that are not representative of the participants' opinions, especially in cases where the participants' opinions are not well-defined or are conflicting.
This can lead to a lack of trust in the generated statements, and it is important to be aware of this limitation when using LLMs in democratic processes.
Finally, the process of generating representative statements is still in its infancy, and there is much room for improvement in terms of the algorithms used and the data collected.
However, most of these limitations can be mitigated by using participatory stages in the process, such as the validation round conducted in the pilot study.
Since LLMs are rapidly-evolving in real time, these issues may not be as concerning in the near future, and it is important to keep an eye on the latest developments in this field.

8 Future Scope

The work done in this report is a step towards the goal of using LLMs in democratic processes, and there is much room for improvement in various aspects.
A good starting point would be to expand the domain of issues at hand. The pilot study in [1] dealt with chatbot personalization, whereas this report deals with educational policies. Other domains too, such as healthcare and climate, could be explored in the future.
Better prompting strategies could also be looked into, in order to tighten the fairness constraints of the statements generated, while maintaining representation of the general views of participants.
Much more advanced voting rules can be a topic of study, such as the use of machine learning algorithms to predict the preferences of participants based on their responses.
This could lead to more accurate and representative statements, and could also help to mitigate the biases and inaccuracies in the generated statements.
Finally, the use of participatory stages in the process could be further explored, in order to increase the transparency and trustworthiness of the generated statements.

9 Conclusion

In the paper "Generative Social Choice" [1] as well as this report, a framework that blends the strengths of generative AI and social choice theory to facilitate representative policy drafting was explored. By collecting free-text responses from participants and leveraging clustering techniques alongside large language models, candidate policy statements were generated that reflected diverse public preferences. To ensure fairness and proportional representation, voting rules such as Balanced Justified Representation (BJR) were applied, which enabled the selection of statements that offer equitable inclusion of justified groups.
The findings demonstrate the feasibility of using LLMs not only to synthesize meaningful content from unstructured input but also to support participatory decision-making through principled selection mechanisms. Importantly, the integration of classical multi-winner voting methods with modern large language models provides a compelling direction for scalable, deliberative systems that maintain democratic values.
This work serves as a step toward more transparent, inclusive, and human-aligned policymaking processes. Future research can explore richer models of preference aggregation, better methods for eliciting values, and further safeguards to ensure factual accuracy and normative neutrality in the generated outputs.

10 Acknowledgements

I would like to thank my Game Theory professors, **Prof. Y Narahari** (narahari@iisc.ac.in) and **Prof. Siddharth Barman** (barman@iisc.ac.in) for giving me the oppurtunity to work on this project, and for their guidance and support throughout the course of this project.
I would also like to thank my project mentors **Y Geetha Charan** and **R K Shishir** for their valuable feedback and suggestions, which have helped me to improve the quality of this report.
I would like to thank all the respondents of the survey for their time and effort in providing their opinions and thoughts on the topic of this project.
Finally, I would like to extend my sincere gratitude to my friends and colleagues, who have supported me throughout this project, and have been a constant source of motivation and encouragement.

References

[1] Sara Fish, Paul Gölz, David C. Parkes, Ariel D. Procaccia, Gili Rusak, Itai Shapira, and Manuel Wüthrich. Generative social choice. *arXiv preprint arXiv:2309.01291*, 2023.

[2] Christopher T. Small, Ivan Vendrov, Esin Durmus, Hadjar Homaei, Elizabeth Barry, Julien Cornebise, Ted Suzman, Deep Ganguli, and Colin Megill. Opportunities and risks of llms for scalable deliberation with polis. *arXiv preprint arXiv:2306.11932*, 2023.

[3] Andrew Konya, Yeping Lina Qiu, Michael P. Varga, and Aviv Ovadya. Elicitation inference optimization for multi-principal-agent alignment, 2022. Manuscript.

[4] Bailey Flanigan, Paul Gölz, Anupam Gupta, Brett Hennig, and Ariel D. Procaccia. Fair algorithms for selecting citizens' assemblies. *Nature*, 596:548–552, 2021.

[5] Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W. Black, and Yulia Tsvetkov. Measuring bias in contextualized word representations. In *Proceedings of the 1st Workshop on Gender Bias in Natural Language Processing*, pages 166–172, 2019.

[6] Min Kyung Lee, Daniel Kusbit, Anson Kahng, Ji Tae Kim, Xinran Yuan, Allissa Chan, Ritesh Noothigattu, Daniel See, Siheon Lee, Christos-Alexandros Psomas, and Ariel D. Procaccia. Webuildai: Participatory framework for fair and efficient algorithmic governance. In *Proceedings of the 22nd ACM Conference on Computer-Supported Cooperative Work and Social Computing (CSCW)*, page Article 181, 2019.

[7] Ritesh Noothigattu, Snehalkumar S. Gaikwad, Edmond Awad, Sohan Dsouza, Iyad Rahwan, Pradeep Ravikumar, and Ariel D. Procaccia. A voting-based system for ethical decision making. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence (AAAI)*, pages 1587–1594, 2018.

[8] Michiel A. Bakker, Martin J. Chadwick, Hannah R. Sheahan, Michael Henry Tessler, Lucy Campbell-Gillingham, Jan Balaguer, Nat McAleese, Amelia Glaese, John Aslanides, Matthew M. Botvinick, and Christopher Summerfield. Fine-tuning language models to find agreement among humans with diverse preferences. In *Proceedings of the 36th Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

[9] Haris Aziz, Markus Brill, Vincent Conitzer, Edith Elkind, Rupert Freeman, and Toby Walsh. Justified representation in approval-based committee voting. *Social Choice and Welfare*, 42(2):461–485, 2017.

[10] Rachel Freedman, Jana Schaich Borg, Walter Sinnott-Armstrong, John P. Dickerson, and Vincent Conitzer. Adapting a kidney exchange algorithm to align with human values. *Artificial Intelligence*, 283, 2020.

[11] Dominik Peters, Grzegorz Pierczynski, and Piotr Skowron. Proportional participatory budgeting with additive utilities. In *Proceedings of the 35th Annual Conference on Neural Information Processing Systems (NeurIPS)*, pages 12726–12737, 2021.

[12] Burt L. Monroe. Fully proportional representation. *American Political Science Review*, 89(4):925–940, 1995.

[13] Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh. Universal adversarial triggers for attacking and analyzing nlp. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019.

[14] Markus Brill and Jannik Peters. Robust and verifiable proportionality axioms for multiwinner voting. In *Proceedings of the 14th ACM Conference on Economics and Computation (EC)*, 2023.

[15] Nick Clegg. Bringing people together to inform decision-making on generative ai. `https://about.fb.com/news/2023/06/generative-ai-community-forum/`, 2023. Blog post.

[16] Wojciech Zaremba, Ara Dhar, Lama Ahmad, Tyna Eloundou, Shibani Santurkar, Sandhini Agarwal, and Jade Leung. Democratic inputs to ai. https://openai.com/blog/democratic-inputs-to-ai, 2023. Blog post.

[17] Daniel Halpern, Gregory Kehne, Ariel D. Procaccia, Jamie Tucker-Foltz, and Manuel Wüthrich. Representation with incomplete votes. In *Proceedings of the 37th AAAI Conference on Artificial Intelligence (AAAI)*, 2023.

[18] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.

[19] Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *arXiv preprint arXiv:2307.03172*, 2023.

[20] AI Institute for Societal Decision Making (AI-SDM). Generative social choice. https://youtu.be/RgmpFUEv3II?si=p8iGZOk7WUg3dFGq, 2024. YouTube video.

Source Code

All the data collected and code pertaining to this project can be found in this GitHub repository.

Appendix

A BJR Algorithm

Algorithm 1 BJR Policy Selection (Concise)

```
procedure BJR(V, k)                          ▷ V: voter preferences, k: number of winners
    n ← number of voters in V
    t ← n/k                                  ▷ Threshold
    S ← empty map from policy to set of voters
    for all v, P in V do
        for all p in P do
            Add v to S[p]
    L ← policies in S sorted by |S[p]| (descending)
    R ← ∅                                    ▷ Selected policies
    C ← ∅                                    ▷ Covered voters
    for all p in L do
        if |S[p]| ≥ t and S[p] ⊈ C then
            Add p to R
            C ← C ∪ S[p]
        if |R| = k then
            break
    return R
```

B Survey Questions

B.1 Generation Survey

Title: Educational Survey
Description: We would like to know your thoughts on the current educational scenario in our country. It would be very helpful if you could spare a few minutes and help us out.

Questions:

1. What do you think are the biggest challenges in the education system today? Please give your thoughts in a sentence or two.
2. How can policymakers ensure that education is accessible to all, regardless of socioeconomic status? Please give your thoughts in a sentence or two.
3. What subjects or skills do you think should be emphasized more in schools? Please give your thoughts in a sentence or two.
4. How should the education system adapt to better prepare students for the modern workforce? Please give your thoughts in a sentence or two.
5. How can universities and industries work together to ensure graduates have relevant job skills? Please give your thoughts in a sentence or two.
6. What role should AI and digital tools play in the future of education? Please give your thoughts in a sentence or two.
7. How can we ensure equal access to technology and the internet for all students? Please give your thoughts in a sentence or two.
8. Do you think online learning should be integrated more into the education system? Why or why not? Please give your thoughts in a sentence or two.

9. What policies could better support teachers and improve their working conditions? Please give your thoughts in a sentence or two.
10. What are your thoughts on the role of private versus public schools in education? Please give your thoughts in a sentence or two.
11. How can parents, teachers, and students have a greater voice in education policy decisions? Please give your thoughts in a sentence or two.
12. What is one change that could significantly improve education in our country? Please give your thoughts in a sentence or two.

B.2 Validation Survey

Title: Statement Ratings
Description: For each statement you read below, kindly rate them on a scale from 1 to 5 based on how much you agree with them, where each digit is as described below.
1 - Not at all 2 - Poorly 3 - Somewhat 4 - Mostly 5 - Perfectly

At the end of each section, please also order the statements in decreasing order of your agreement with them.

Section 1:

S1 India's education policy should focus on elevating public schools through better infrastructure, trained and fairly compensated teachers, inclusive decision-making, and updated curricula that emphasize critical thinking, real-world skills, and digital access—while fostering healthy collaboration and accountability between public and private institutions to ensure equitable and high-quality education for all.

S2 India's education policy should focus on improving the quality and accessibility of public education by investing in infrastructure, technology, teacher training and compensation, curriculum updates emphasizing critical thinking, problem-solving, and relevant skills, while also establishing better oversight and accountability for both public and private institutions through feedback mechanisms and potentially public-private partnerships to ensure equitable and quality education for all.

S3 India's education policy should focus on improving the quality of public schools through better teacher compensation, enhanced infrastructure, critical thinking-centered curricula, and digital literacy, while establishing stronger accountability mechanisms that incorporate feedback from teachers, students, and parents to create an accessible and forward-looking educational system that provides opportunities for all children regardless of socioeconomic background.

Please rate the above statements in decreasing order of your agreement with their contents.

Section 2:

S1 India's education policy should focus on enhanced collaboration with industry to integrate practical skills, provide real-world exposure through internships and project-based learning, and ensure the curriculum remains relevant and equips graduates with adaptable skills for the job market.

S2 India's education policy should focus on fostering strong collaboration between academia and industry to promote hands-on, project-based learning and develop adaptable, real-world skills such as communication, critical thinking, and digital literacy.

S3 India's education policy should focus on strengthening collaboration between academic institutions and industry through internship programs, project-based learning, and co-designed curricula that balance theoretical knowledge with practical skills development, ensuring students graduate with relevant competencies for the modern workforce.

Please rate the above statements in decreasing order of your agreement with their contents.

Section 3:

S1 India's education policy should focus on increasing awareness through internet-accessible resources like standard-based video lectures, providing better guidance for students, developing meaningful project-based learning, and establishing a quality oversight board to improve underperforming institutions.

S2 India's education policy should focus on increasing awareness, ensuring access to quality standardized resources like government-approved video lectures, promoting meaningful project-based learning, and establishing a transparent body to uplift or reform substandard institutions.

S3 India's education policy should focus on strategically integrating AI and digital tools to personalize learning, enhance teaching effectiveness, and provide flexibility, while also prioritizing teacher quality through better compensation, rigorous selection, and continuous training, and shifting towards more practical, competency-based assessments that emphasize completion over competition.

Please rate the above statements in decreasing order of your agreement with their contents.

Section 4:

S1 India's education policy should focus on revamping primary education to prioritize core subjects like mathematics, science, and economics, while eliminating outdated content and ensuring better teaching quality.

S2 India's education policy should focus on increasing awareness and guidance, potentially through accessible resources like cheaper internet, awareness campaigns, recorded standard lectures, government-approved video content, and establishing a quality board to address underperforming institutions and ensure standardized education.

S3 India's education policy should focus on goal-oriented learning that aligns curriculum with socioeconomic needs, emphasizes practical outcomes for employment, and adapts regularly to changing job market requirements, all while fostering student purpose and self-care.

Please rate the above statements in decreasing order of your agreement with their contents.

C Prompts for Statement Generation

The same prompt was used on all three LLMs, to ensure consistency.

LLM Prompt

Given the following responses from a survey on Indian Education Policy, generate a single sentence that represents their common viewpoint. Start the response with "India's education policy should focus on":
`$Responses_i`

Here, the placeholder `$Responses_i` is replaced with the responses from each cluster, to be used as a means of multi-shot prompting.