Generating Fair Consensus Statements with Social Choice on Token-Level MDPs*

Carter Blair
Cheriton School of Computer Science
University of Waterloo
Waterloo, Canada
cblair@uwaterloo.ca

Kate Larson
Cheriton School of Computer Science
University of Waterloo
Waterloo, Canada

Abstract

Current frameworks for consensus statement generation with large language models lack the inherent structure needed to provide provable fairness guarantees when aggregating diverse free-form opinions. We model the task as a multi-objective, token-level Markov Decision Process (MDP), where each objective corresponds to an agent's preference. Token-level rewards for each agent are derived from their policy (e.g., a personalized language model). This approach utilizes the finding from Rafailov et al. [17] that such policies implicitly define optimal Q-functions, providing a principled way to quantify rewards at each generation step without a value function. This MDP formulation creates a formal structure amenable to analysis using principles from social choice theory. We propose two approaches grounded in social choice theory. First, we propose a stochastic generation policy guaranteed to be in the ex-ante core, extending core stability concepts from cooperative game theory and voting theory to text generation. This policy is derived from an underlying distribution over complete statements that maximizes proportional fairness (Nash Welfare). Second, for generating a single statement, we target the maximization of egalitarian welfare using search algorithms within the MDP framework. Empirically, we find that search guided by the egalitarian objective generates consensus statements with improved worst-case agent alignment compared to baseline methods, including the Habermas Machine [22].

1 Introduction

Social choice theory has traditionally addressed the aggregation of preferences over predefined sets of alternatives. Large Language Models (LLMs) now enable the aggregation of free-form, verbal opinions into collective textual outputs, offering greater flexibility. A key benefit of aggregating free-form opinions is the potential to reduce the agenda-setting power often held by organizers of collective decisions, as participants are not constrained to predefined choices. However, ensuring provable fairness in these outputs presents a challenge, particularly as fairness itself can be conceptualized in many ways, and the intricate, high-dimensional computations within LLMs make auditing for specific fairness criteria throughout the generation process challenging. For this reason, previous approaches have treated the generation process as a black box, applying various fairness measures post-hoc. For instance, in the Habermas Machine [22], statements are generated, and fairness is then pursued through a voting procedure applied to these statements, which were not themselves generated with an explicitly fair mechanism. Similarly, one part of the Generative Social Choice method [10] attempts to maximize egalitarian welfare by prompting an LLM to do so. These strategies, while aiming for fairness, can inadvertently cede a new form of agenda control to the LLM itself. By tasking the LLM

*Our code is available on GitHub.

Preprint. Under review.

to directly produce an output that satisfies a broad fairness objective or goal, the specific ways the LLM interprets and operationalizes this directive, the implicit trade-offs it makes, or the aspects of opinions it prioritizes remain opaque. This effectively allows the LLM to shape the solution space. As such, these methods risk overlooking biases embedded within the generation process itself, which are known to exist [9].

We address this gap by modeling consensus statement generation as a token-level Markov Decision Process (MDP). Each agent i's viewpoint is represented by a policy π_i, which assigns likelihoods $\pi_i(s, a)$ to token choices given the current prefix s. Following Rafailov et al. [17], who show that policies implicitly define optimal Q-functions, our agent policies π_i determine rewards $r_i(s, a)$ (e.g., $r_i^{\log}(s, a) = \beta \log \pi_i(s, a)$) at each generation step. A primary advantage of this reward formulation is that it avoids personalized value functions, which are known to be challenging to train and apply effectively [12]. This MDP structure provides a formal basis for integrating fairness principles directly into the construction of the consensus statement.

Within this MDP framework, we develop two approaches that leverage existing notions of fairness from social choice theory, namely the *ex-ante core* and *egalitarian welfare (EW)*, which we argue are compelling notions of fairness in the context of consensus statement generation. First, to achieve an outcome in the ex-ante core, we propose a stochastic generation policy Π^*. This policy is derived by optimizing a distribution over complete statements to maximize proportional fairness (Nash Welfare), a process known to yield core membership. For consensus generation, the core is a highly desirable stability concept: a lottery in the core ensures that no coalition of agents could unilaterally deviate and achieve an alternative lottery that all its members prefer, given their proportional influence, implying agreement, nay consensus, over the randomized outcome. Second, when a single consensus statement is desired, we target the maximization of EW, seeking the best outcome for the least satisfied agent, which aligns with the idea that a consensus statement should be agreeable to all parties. We introduce constructive search algorithms (finite lookahead and beam search) that optimize this EW objective directly within the MDP. This offers a transparent and analyzable generation mechanism distinct from methods reliant on high-level prompting or post-hoc voting.

Our main contributions are:

1. A formal token-level MDP framework for fair consensus generation where agent rewards are derived from their language model policies.
2. A method to derive a stochastic generation policy that is provably in the ex-ante core, ensuring proportional fairness and stability.
3. The development and empirical validation of search algorithms that, by optimizing egalitarian welfare within the MDP, generate single consensus statements with improved worst-case agent alignment compared to methods that do not leverage this token-level structure or search.

Through these contributions, we hope to establish a new direction for methods seeking to generate consensus statements from open-ended verbal opinions with provable fairness guarantees.

2 Related Work

Generative Social Choice. This field applies social choice principles to open-ended generation, such as creating text from diverse opinions [10, 19, 22]. Unlike methods that aggregate preferences over predefined alternatives, Generative Social Choice (GSC) generates the alternatives themselves. For example, Tessler et al. [22] employ iterative critiques and voting on complete statements for consensus, while Fish et al. [10] prompt LLMs to directly maximize egalitarian welfare. Our work differs by embedding fairness into the token-by-token construction of a consensus statement via a multi-objective MDP, treating each token selection as a public decision [4]. This provides a more granular and verifiable mechanism than post-hoc evaluations or high-level prompting.

Randomized Social Choice. We also draw from randomized social choice, which studies lotteries over outcomes. Our stochastic generation policy, which maximizes Nash Welfare for proportional fairness, connects to this area. Maximizing Nash Welfare is known to yield outcomes in the 1-core [1, 7, 8]. We extend these findings to the sequential decision-making context of this paper.

Guided Decoding. Guided decoding techniques steer LLM generation towards desired attributes at inference time, often using search algorithms. Methods like PPO-MCTS [14] and VAS [13] use a value network to guide generation, while MOD [21] or COLLAB [3] combine or switch policies. Our approach also uses search but derives token-level rewards from agent policies. Further, we explicitly frame generation as planning in a multi-objective MDP to optimize social choice objectives (Proportional Fairness, Egalitarian Welfare), rather than relying on a single pre-trained value model or heuristic model combinations.

3 Problem Setup & Preliminaries

Consider a setting with a finite set of agents $N = \{1, 2, \ldots, n\}$, each with a distinct opinion on a specific **Issue**. The goal is to generate a consensus text statement reflecting these perspectives fairly. The inputs to the process include descriptions of the Issue, the **agent opinions** (e.g., free-form text expressing their views), and the derived **agent policies**. Additionally, a **reference consensus policy** (e.g., a base language model) is used to propose tokens in the consensus statement.

Agent Policies. Each agent $i \in N$ is represented by a *policy* π_i, which assigns a likelihood $\pi_i(s, a) \in [0, 1]$ to each action a given the state s. Intuitively, $\pi_i(s, a)$ reflects how closely an action aligns with agent i's preference at state s. This policy could be an LLM fine-tuned (ideally with DPO [16]) or prompted for agent i's viewpoint.

Token-Level MDP. We model text generation as a deterministic, discrete-time Markov Decision Process defined by the tuple $(S, A, T, \mathbf{R})$. Here, S is the state space of partial text sequences (prefixes), including initial s_0 and terminal states. A is the action space consisting of the token vocabulary plus a special end-of-sequence token $\langle\text{eos}\rangle$. T is the deterministic transition function where $T(s, a) = s\|a$ appends the chosen token; selecting $a = \langle\text{eos}\rangle$ leads to a terminal state representing a completed statement X. Finally, $\mathbf{R}$ represents the agent-specific reward functions. We define two types of rewards based on agent policies, serving different analytical purposes:

1. **Log-Likelihood Reward:** $r_i^{\log}(s, a) = \beta \log \pi_i(s, a)$. This formulation aligns with implicit rewards in preference learning [17], where $\beta > 0$ is a scaling factor. This is non-positive and is suitable for additive utility accumulation along a path.
2. **Likelihood Reward:** $r_i^{\text{prob}}(s, a) = \pi_i(s, a)$. This reward uses the direct probability, ensuring non-negativity ($r_i^{\text{prob}} \geq 0$), which is needed for social welfare functions involving products or ratios, such as Nash Welfare.

We denote by $\mathcal{C}$ the set of all possible complete paths (sequences ending in $\langle\text{eos}\rangle$) from s_0.

In practice, we assume that at each non-terminal state s, we only consider a finite set of B possible next tokens $A_B(s) \subseteq A$. This set could be the model's vocabulary or a subset chosen by a base language model ranking tokens by probability. With this finite branching factor B and a maximum sequence length $L_{\max}$, the set $\mathcal{C}$ of complete paths is finite (bounded by $B^{L_{\max}}$).

This sequential token selection process naturally defines a tree structure rooted at s_0. Each edge represents choosing a token from $A_B(s_t)$, and each node represents a partial sequence s_t. To make this concrete, we illustrate an example in Figure 1.

Agent Utilities. Given a completed sequence $X = (a_1, \ldots, a_\ell = \langle\text{eos}\rangle)$ corresponding to states $(s_0, s_1, \ldots, s_\ell)$, we define two corresponding utility functions for each agent i, derived from the respective reward types:

1. **Additive Log-Utility:** Primarily used for evaluating single paths based on cumulative log-likelihood.

$$U_i^{\log}(X) = \sum_{t=1}^{\ell} r_i^{\log}(s_{t-1}, a_t) = \sum_{t=1}^{\ell} \beta \log \pi_i(s_{t-1}, a_t) = \beta \log \left(\prod_{t=1}^{\ell} \pi_i(s_{t-1}, a_t) \right)$$

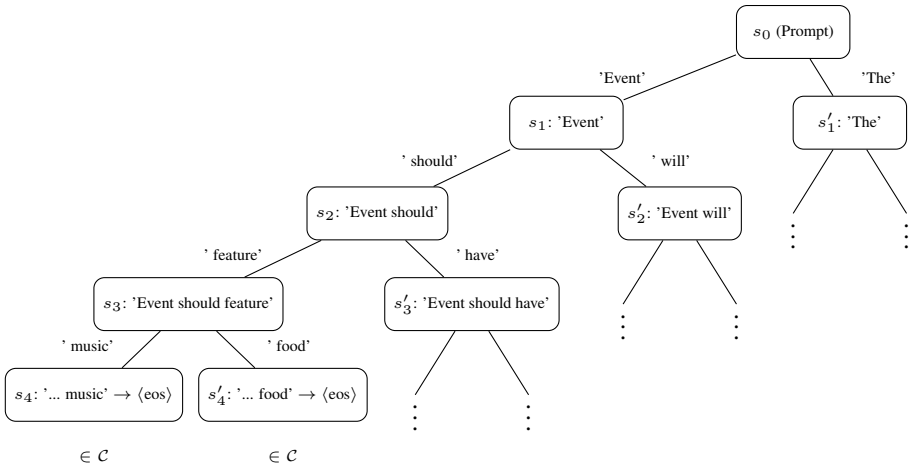

Figure 1: Illustration of the token-level generation tree. Edges represent actions (chosen tokens).

2. **Multiplicative Probability Utility:** Primarily used for evaluating distributions via expected utility, forming the basis for Nash Welfare and Proportional Fairness calculations.

$$U_i^{\text{prob}}(X) = \prod_{t=1}^{\ell} r_i^{\text{prob}}(s_{t-1}, a_t) = \prod_{t=1}^{\ell} \pi_i(s_{t-1}, a_t) = P_i(X)$$

This represents the joint probability of sequence X under agent i's policy.

These are related by $U_i^{\log}(X) = \beta \log U_i^{\text{prob}}(X)$, with $U_i^{\text{prob}}(X) > 0$.

A Single Path vs. a Lottery. We consider two types of outcomes:

1. A single deterministic path $X \in \mathcal{C}$. For single paths, we adopt additive log-utilities $U_i^{\log}(X)$.
2. A distribution $p \in \Delta(\mathcal{C})$ over paths (a lottery) which can induce a stochastic policy. When assessing this type of outcome, we consider the expected probability-based utility.

$$\mathbb{E}_{X\sim p}[U_i^{\text{prob}}(X)] = \sum_{X\in\mathcal{C}} p(X) U_i^{\text{prob}}(X).$$

When it is clear from the context, we adopt the shorthand of $U_i^{\text{prob}}(p)$ to refer to the expected utility of a distribution. This expected utility $U_i^{\text{prob}}(p)$ is guaranteed to be non-negative, satisfying the requirements of downstream fairness measures.

The Fairness of a Path. For a single generated path $X \in \mathcal{C}$, we assess fairness by its egalitarian welfare (EW), drawing from Rawls' maximin principle [18]. This is defined using the additive log-utilities $U_i^{\log}(X)$:

$$\text{EW}^{\log}(X) = \min_{i\in N} U_i^{\log}(X) = \min_{i\in N} \sum_{t=1}^{\ell} \beta \log \pi_i(s_{t-1}, a_t). \tag{1}$$

Maximizing $\text{EW}^{\log}(X)$ means finding the path whose cumulative log-likelihood is highest for the agent who values it least. This approach aims to maximize the well-being of the least satisfied agent, thereby reducing the risk of marginalizing minority viewpoints or producing a statement unacceptable to some participants. By focusing on the minimum utility, the egalitarian objective promotes broadly acceptable outcomes, fostering inclusivity and supporting the requirement that a consensus statement be agreeable to all parties.

The Fairness of a Lottery. To analyze the fairness of lotteries (distributions $p \in \Delta(\mathcal{C})$ over paths), we use criteria based on non-negative expected utilities $U_i^{\text{prob}}(p)$. The first criterion is proportional fairness (PF), achieved by maximizing the Nash social welfare (NW) function [7]:

$$\text{NW}(p) = \prod_{i=1}^{n} U_i^{\text{prob}}(p).$$

Maximizing $NW(p)$ yields a proportionally fair distribution [7].

The second criterion is the *core* [20], a concept from cooperative game theory that has been adapted to randomized voting. An outcome in the core is stable, meaning no coalition of agents can unilaterally achieve a better outcome for all of its members, considering their proportional influence [1]. This resistance to coalitional deviation is desirable for consensus statements, as a lottery in the core represents an outcome that all parties have implicitly agreed to, as no group (or single agent) could construct a preferred alternative distribution with their share of probability mass. A distribution maximizing Nash Welfare (using U_i^{prob}) is guaranteed to be in the 1-core [1, 7, 8].

Definition 1 (α-Core)**.** *For $\alpha \geq 1$, a distribution $p \in \Delta(\mathcal{C})$ is in the α-core if there is no coalition $S \subseteq N$ and alternative distribution p' such that*

$$\frac{|S|}{|N|} \cdot U_i^{prob}(p') \geq \alpha \cdot U_i^{prob}(p), \quad \forall i \in S,$$

with strict inequality for at least one agent $i \in S$. When $\alpha = 1$, this is also simply referred to as the core.

4 Stochastic Policies and Ex-Ante Fairness

Having established the token-level MDP and fairness criteria, we now turn to defining a *stochastic generation policy* Π^* that produces a distribution p_{Π^*} over complete consensus statements $\mathcal{C}$ satisfying ex-ante fairness properties. Specifically, we aim for the generated distribution to be **proportionally fair** and reside in the **1-core**. Our strategy is to first identify the optimal target distribution p^* and second, derive a sequential policy Π^* that, when executed during generation, yields exactly p^*.

The potentially enormous size of the set of all possible statements $\mathcal{C}$, which can grow exponentially with the maximum sequence length $L_{\max}$ and branching factor B, presents a computational challenge. Directly computing p^* over the full space $\mathcal{C}$ could be computationally intractable in large trees.

4.1 Computational Tractability via Token Chunking

To address the computational burden, we introduce a *chunking* strategy. This approach groups sequences of tokens into larger units, reducing the depth and branching factor of the decision tree.

Definition 2 (Token Chunking)**.** *A chunking strategy $\mathcal{K}$ partitions the token sequence into contiguous segments (chunks) $\{k_1, k_2, \ldots, k_m\}$. Each chunk k_j consists of one or more tokens. Their concatenation forms a complete statement. Actions now correspond to selecting entire chunks.*

A simple fixed-size chunking strategy where each chunk has size c (a hyperparameter) reduces the effective path length from $L_{\max}$ tokens to $\lceil L_{\max}/c \rceil$ chunks. The search space is thus restricted to $\mathcal{C}_\mathcal{K}$, the set of complete paths constructible using these chunks.

This chunking improves computational feasibility but introduces an approximation: the optimal proportionally fair lottery over the full space $\mathcal{C}$ might involve paths not representable in the chunked space $\mathcal{C}_\mathcal{K}$. Our method finds the lottery $p^* \in \Delta(\mathcal{C}_\mathcal{K})$ that maximizes Nash Welfare *relative to this restricted space*. Consequently, the fairness guarantees (proportional fairness, 1-core membership) hold within $\mathcal{C}_\mathcal{K}$, but the expected utilities achieved might be suboptimal compared to what was possible in the unchunked space $\mathcal{C}$. The choice of chunk size c thus involves a trade-off between tractability and the potential optimality gap.

4.2 Deriving the Ex-Ante Fair Stochastic Policy

Given the target distribution $p^* \in \Delta(\mathcal{C}_\mathcal{K})$ that maximizes Nash Welfare over the chunked space, we now derive the stochastic policy Π^* that generates this distribution. Executing Π^* involves sequentially sampling the next chunk based on probabilities derived from p^*.

To define Π^*, we introduce some notation. For any state (prefix) s in the chunked generation tree:

- Let $\mathcal{C}_\mathcal{K}(s) \subseteq \mathcal{C}_\mathcal{K}$ be the set of all complete paths (leaves) in the chunked space that pass through state s.
- Let $\mathcal{C}_\mathcal{K}(s,k) \subseteq \mathcal{C}_\mathcal{K}(s)$ be the subset of paths in $\mathcal{C}_\mathcal{K}(s)$ where the next action (chunk) taken from state s is k. Note that $\mathcal{C}_\mathcal{K}(s,k) = \mathcal{C}_\mathcal{K}(s\|k)$, where $s\|k$ is the state reached after taking chunk k.
- For any subset of leaves $L \subseteq \mathcal{C}_\mathcal{K}$, let $P^*(L) = \sum_{X \in L} p^*(X)$ be the total probability mass assigned by the optimal lottery p^* to the leaves in L.

Note that $P^*(\mathcal{C}_\mathcal{K}(s_0)) = P^*(\mathcal{C}_\mathcal{K}) = 1$, where s_0 is the initial empty state.

With this notation, we can define the policy Π^* at any given state s.

Definition 3 (Stochastic Policy Induced by Lottery p^*)**.** *Let p^* be a distribution over the leaf nodes $\mathcal{C}_\mathcal{K}$. The induced stochastic policy Π^* at a non-terminal state s assigns the probability of taking the next action (chunk) k as:*

$$\Pi^*(s,k) = \begin{cases} \frac{P^*(\mathcal{C}_\mathcal{K}(s,k))}{P^*(\mathcal{C}_\mathcal{K}(s))} & \text{if } P^*(\mathcal{C}_\mathcal{K}(s)) > 0 \\ 0 & \text{if } P^*(\mathcal{C}_\mathcal{K}(s)) = 0 \end{cases} \tag{2}$$

This represents the conditional probability, according to the target distribution p^*, of selecting chunk k next, given that the generation process has reached state s. If state s has zero probability of being reached under p^* (i.e., $P^*(\mathcal{C}_\mathcal{K}(s)) = 0$), then the probability of taking any action from s is also zero.

4.3 Properties of the Induced Policy

We now establish that executing this policy Π^* from the initial state s_0 indeed generates the target distribution p^* over the leaves of the chunked tree $\mathcal{C}_\mathcal{K}$.

Theorem 1 (Equivalence of Policy-Induced Distribution and Target Lottery)**.** *Let p_{Π^*} be the distribution over $\mathcal{C}_\mathcal{K}$ generated by executing the policy Π^* (defined in Definition 3) from the initial state s_0. Then $p_{\Pi^*} = p^*$.*

The proof is presented in Appendix C. This equivalence directly leads to the desired ex-ante fairness guarantee for the policy Π^*.

Corollary 1 (Core Membership of Stochastic Policy)**.** *Let p^* be a distribution over $\mathcal{C}_\mathcal{K}$ that maximizes Nash Welfare (and is therefore in the 1-core relative to $\mathcal{C}_\mathcal{K}$). Let Π^* be the stochastic policy derived from p^* according to Definition 3. Then the distribution p_{Π^*} generated by executing Π^* is in the 1-core relative to $\mathcal{C}_\mathcal{K}$.*

Proof. By Theorem 1, the distribution generated by policy Π^* is $p_{\Pi^*} = p^*$. Since p^* was chosen to maximize Nash Welfare over $\mathcal{C}_\mathcal{K}$, it is in the 1-core relative to this set. Therefore, p_{Π^*} is also in the 1-core relative to $\mathcal{C}_\mathcal{K}$. □

This result confirms that our procedure, which first finds the distribution maximizing Nash Welfare p^* over the chunked space $\mathcal{C}_\mathcal{K}$ and then executes the derived policy Pi^*, yields a stochastic policy in the core.

5 Generating a Single Consensus Statement

While our stochastic policy Π^* offers strong ex-ante fairness guarantees, some practical applications require selecting a single consensus statement. In this case, our objective shifts from finding a fair distribution to identifying the single path that best represents all agents' preferences.

5.1 Finding the Rawlsian Path

Given the token tree with leaf nodes $\mathcal{C}$ (potentially derived from chunking, $\mathcal{C}_\mathcal{K}$), we aim to find a single path $X^* \in \mathcal{C}$ that maximizes egalitarian welfare as defined in Equation 1. Due to the size of the token tree, exhaustive search for X^* may be intractable. We propose approximate algorithms to find high-quality paths, including finite-lookahead search and beam search, which are detailed below.

Finite Lookahead Search. The finite lookahead algorithm operates with a rolling horizon. At each step t, it explores all possible paths P of length up to d originating from s_t. For each such path P, the algorithm evaluates the egalitarian welfare of the sequence formed by concatenating the path generated so far (X_{prefix}) with P. It then chooses the first action a^* of the path P^* that maximizes this lookahead evaluation, transitions to state $s_{t+1} = T(s_t, a^*)$, and repeats the process. This d-step lookahead can mitigate the potential for hedging inherent in greedy search. When no single immediate token is agreeable (i.e., results in high egalitarian welfare), a greedy method might select less informative tokens that avoid commitment. In contrast, a lookahead can identify longer sequences that, despite potentially controversial initial steps, lead to states with higher overall welfare, perhaps by expressing a concept with suitable qualifications. The algorithm is shown in Appendix D.

Beam Search. Beam search is a heuristic search algorithm that balances greedy search and exhaustive exploration, proving effective in sequence generation tasks like machine translation and text generation [15, 17]. Instead of pursuing only the single best option (greedy search) or all options (exhaustive search), beam search maintains a fixed number of the most promising partial paths (hypotheses), w (the beam width), at each depth t. At each step, it expands paths in the beam by generating potential successor tokens. These candidates are then evaluated using the egalitarian welfare objective function, and only the top w scoring paths are retained for the next step. This approach allows beam search to explore a more diverse set of sequences than greedy search, mitigating the risk of suboptimal paths from poor early choices, while remaining computationally tractable. The algorithm returns the highest-scoring complete path found within the beam at the maximum length or upon reaching a terminal state. The algorithm is shown in Appendix D.

6 Experiments

We conduct experiments to, first, test whether agent policies derived from prompting language models can perform credit-assignment to provide meaningful token-level rewards, which is a prerequisite for our search algorithms. Second, we assess the performance of the finite lookahead and beam search methods in generating single consensus statements compared to baseline approaches. All of our experiments were run on cloud servers and can be run in approximately a week with 5 CPUs.

6.1 Evaluating Prompt-Based Credit Assignment

Our framework defines token-level rewards $r_i(s, a) = \beta \log \pi_i(s, a)$ using policies π_i derived from prompting a base LLM with agent-specific information. Effective reward-guided search requires these policies to exhibit "localized credit assignment," where policy likelihoods $\pi_i(s, a)$ change primarily at tokens relevant to the prompt's information. While observed in DPO-trained models [17], we empirically validate this for policies derived from prompted instruction-tuned models.

We conducted controlled experiments with `Llama 3.1 8B Instruction-Tuned` and `Gemma 2 9b Instruction-Tuned` to measure if profile information localizes its influence on token likelihoods. For each test, we compared token log-probabilities from a **user policy prompt** (e.g., "User time profile: morning") against a **reference policy prompt** (e.g., "User time profile: empty"). This comparison was done for two nearly identical sequences: X_1, with a concept conflicting with the user policy (e.g., "spaghetti" for a "morning" profile), and X_2, where the conflicting concept was replaced with an aligned one (e.g., "pancakes").

To quantify localized effects, for each token a_j and its preceding sequence s, we calculated the difference in log-likelihood under the user policy (π_U) versus the reference policy (π_R): $\Delta L(a_j|s) = \log \pi_U(a_j|s) - \log \pi_R(a_j|s)$. We then measured the absolute change in this $\Delta L(a_j|s)$ when switching between sequences X_1 and X_2: $|\Delta L_{X_1}(a_j|s) - \Delta L_{X_2}(a_j|s)|$. A large value for this absolute change indicates the profile's influence on that token a_j shifted substantially with the aligned/misaligned concept swap. Finally, we calculated a Z-score for each token j based on this shift, relative to the mean and standard deviation of shifts across all tokens in the sequence. A high Z-score highlights tokens where the user policy prompt induced a statistically significant preference shift when the input concept was swapped, pinpointing localized credit assignment.

As hypothesized, the largest Z-scores occurred at the key tokens that varied between X_1 and X_2 (Table 1). For example, with a "morning" time profile, the "spaghetti/pancakes" token pair showed the most significant shift. This, along with the additional examples with Llama and `Gemma 2 9b`

`Instruction-Tuned` in Subsection E.1, confirms that system prompting induces localized credit assignment in instruction-tuned models, validating that $r_i(s, a) = \beta \log \pi_i(s, a)$ from our prompted policies provides targeted signals for token-level search.

Table 1: Example of credit assignment for `Llama 3.1 8B Instruction-Tuned`. Darker green indicates larger Z-score. Z-score column is for altered tokens (a_j).

Reference policy prompt	**User policy prompt**	**Sequence**	**Z-Score**
User time profile: empty	User time profile: morning	I am about to eat some food . I am going to have spaghetti/pancakes . I will use my phone to order it .	4.26

6.2 Consensus Generation

We evaluated different approaches for generating a single consensus statement by comparing our proposed search algorithms against several baselines. The primary goal was to assess how well each method optimizes egalitarian welfare (EW), measured by a perplexity-based metric reflecting worst-case agent alignment. More detailed consensus generation experiments, with `Gemma 2 9b Instruction-Tuned` and an LLM-judge metric are presented in Subsection E.4 of Appendix E.

6.2.1 Experimental Setup

Scenarios: We used scenarios from the Habermas Machine dataset [22]. To obtain distinct settings, scenario descriptions were embedded using `BAAI/bge-large-en-v1.5` [23] and clustered via k-means ($k = 3$). Representative scenarios were selected from each cluster (Scenarios 1, 2, and 3). The issues for these scenarios are in the captions of Tables 5, 6, and 7.

Agents and Policies: For each scenario, agent opinions were taken from the dataset. Agent policies π_i were instantiated by prompting `Llama 3.1 8B Instruct` [11] with the issue and agent i's opinion, instructing it to generate text aligned with that viewpoint (prompt shown in Figure 5 in Appendix F). The resulting likelihoods $\pi_i(s, a)$ represent agent i's preferences. Agent opinions are detailed in Tables 5, 6, and 7 in Subsection E.3 of Appendix E.

Base Generation Model: Consensus statements were generated using `Llama 3.1 8B Instruct`, prompted with the issue and all agent opinions (prompt shown in Figure 4 in Appendix F).

Evaluation Metric - Egalitarian Perplexity (EPPL): To capture alignment with the least satisfied agent for a consensus statement X, we define Egalitarian Perplexity. For each agent i, their specific perplexity $PPL_i(X)$ is found by prompting `Llama 3.1 8B Instruct` with the issue and agent i's opinion to generate a statement perfectly reflecting that opinion. The average log-likelihood of the actual consensus statement $X = (a_1, \ldots, a_L)$ conditioned on this agent-specific prompt is:

$$\bar{L}_i(X) = \frac{1}{L} \sum_{t=1}^{L} \log \pi_i(s_{t-1}, a_t | \text{prompt}_i).$$

The agent-specific perplexity is $PPL_i(X) = \exp(-\bar{L}_i(X))$. The final Egalitarian Perplexity for X is $EPPL(X) = \max_{i \in N} PPL_i(X)$. Lower EPPL indicates better egalitarian welfare.

Seeds: We report the mean and standard deviation of EPPL over 3 seeds per method and scenario.

Methods Compared: We compared our proposed algorithms, **Finite Lookahead** (Algorithm 1, depth $d = 4$, branching $B = 2$), and **Beam Search** (Algorithm 2, width $w = 4$, pruning based on partial EPPL), against three baselines: **Best-of-N** (selecting the best of $N = 4$ samples[2] frxom the base model by EPPL); a **Prompted Habermas Machine**[3] (1 critique round, 4 candidates[4], critiques from base model conditioned on agent opinions); and the original **Habermas Machine** (HM) [22] consensus (generated by a fine-tuned Chinchilla-70B).

[2] $N = 4$ was chosen to align with the Prompted Habermas Machine.

[3] As implemented in https://github.com/google-deepmind/habermas_machine

[4] We chose four candidates to align with the default parameters in the Prompted Habermas Machine example in the Habermas Machine GitHub repository.

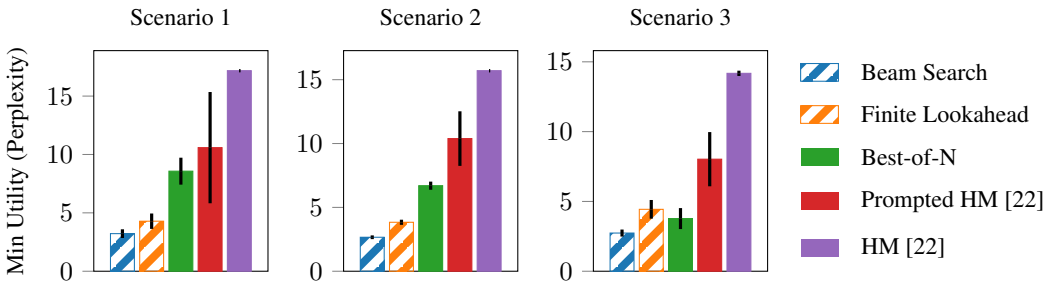

Figure 2: Per-scenario egalitarian welfare (perplexity). Lower values indicate better minimum agent utility. Striped bars indicate that the method uses search over the token-level MDP. Numerical results are reported in Table 4 in Appendix E.

6.2.2 Results

Figure 2 summarizess EPPL performance (lower is better). **Beam Search consistently achieved the lowest EPPL** (average: 2.87), indicating high alignment with the least satisfied agent. **Finite Lookahead also performed well** (average EPPL: 4.18), outperforming baseline methods. Both search methods surpassed **Best-of-N** (6.35) and the **Prompted Habermas Machine** (9.67). The **Habermas Machine** baseline had the highest EPPL (15.69), likely because its statement was generated by a different model (Chinchilla 70B).

The results suggest that token-level search guided by EPPL, as in Beam Search and Finite Lookahead, effectively generates consensus statements with better minimum agent alignment compared to sampling or iterative refinement. Consensus statements for the first seed are in Tables 5, 6, and 7. The strong empirical performance of methods operating on the token-level MDP complements the fact that these methods are also more amenable to theoretical analysis and fairness guarantees.

7 Discussion

This work introduced a framework for generating consensus statements by modeling the process as a multi-objective, token-level MDP with rewards derived from agent-specific language model policies. Our aim was to connect LLM-based text generation with the formal fairness guarantees of social choice theory via this MDP.

Our theoretical contributions for stochastic outcomes (lotteries over statements) focused on proportional fairness (PF) and the core. By maximizing Nash Welfare over expected probability-based utilities, we identified an optimal lottery p^* that induces a stochastic generation policy Π^* (Definition 3) inheriting the 1-core property (Corollary 1). Chunking was introduced as a heuristic to manage the search space. For deterministic outcomes (single statements), we focused on maximizing egalitarian welfare (EW), proposing finite lookahead and beam search as approximation algorithms.

Empirical results validated several aspects of our framework. Credit assignment experiments (Subsection 6.1) confirmed that prompting LLMs with agent profiles creates policies π_i whose token likelihoods reflect specified viewpoints, supporting $r_i(s, a) = \beta \log \pi_i(s, a)$ as a meaningful reward. Consensus generation experiments (Subsection 6.2), using Egalitarian Perplexity (EPPL) to measure EW, showed that beam search and finite lookahead, guided by the EW objective, outperformed baselines like Best-of-N and an adapted Habermas Machine. Beam search yielded the lowest EPPL.

Overall, formulating consensus generation as a search problem within a token-level MDP, guided by explicit social choice objectives like EW, is a promising direction. Search-based methods demonstrated advantages in optimizing minimum agent utility. However, we note that what we mean by "utility" is mathematical and, while in theory it could connect to a human's true utility [17], it is not guaranteed. Human studies should evaluate the degree to which these notions correspond. Further, we note that until our methods are better understood, the outputs of our algorithm should be treated as artifacts for collective sense-making instead of binding decisions, as suggested by Revel and Pénigaud [19]. In sum, this work contributes theoretical foundations and practical algorithms for incorporating social choice principles into generative AI for collective decision-making and sense-making.

References

[1] Haris Aziz, Anna Bogomolnaia, and Hervé Moulin. Fair mixing: the case of dichotomous preferences. In *Proceedings of the 2019 ACM Conference on Economics and Computation*, pages 753–781, 2019.

[2] Ahmad Beirami, Alekh Agarwal, Jonathan Berant, Alexander D'Amour, Jacob Eisenstein, Chirag Nagpal, and Ananda Theertha Suresh. Theoretical guarantees on the best-of-n alignment policy. *arXiv preprint arXiv:2401.01879*, 2024.

[3] Souradip Chakraborty, Sujay Bhatt, Udari Madhushani Sehwag, Soumya Suvra Ghosal, Jiahao Qiu, Mengdi Wang, Dinesh Manocha, Furong Huang, Alec Koppel, and Sumitra Ganesh. Collab: Controlled decoding using mixture of agents for LLM alignment. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=7ohlQUbTpp.

[4] Vincent Conitzer, Rupert Freeman, and Nisarg Shah. Fair public decision making. In *Proceedings of the 2017 ACM Conference on Economics and Computation*, pages 629–646, 2017.

[5] Avinava Dubey, Zhe Feng, Rahul Kidambi, Aranyak Mehta, and Di Wang. Auctions with llm summaries. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 713–722, 2024.

[6] Paul Duetting, Vahab Mirrokni, Renato Paes Leme, Haifeng Xu, and Song Zuo. Mechanism design for large language models. In *Proceedings of the ACM Web Conference 2024*, pages 144–155, 2024.

[7] Soroush Ebadian, Anson Kahng, Dominik Peters, and Nisarg Shah. Optimized distortion and proportional fairness in voting. *ACM Transactions on Economics and Computation*, 12(1):1–39, 2024.

[8] Brandon Fain, Kamesh Munagala, and Nisarg Shah. Fair allocation of indivisible public goods. In *Proceedings of the 2018 ACM Conference on Economics and Computation*, pages 575–592, 2018.

[9] Shangbin Feng, Chan Young Park, Yuhan Liu, and Yulia Tsvetkov. From pretraining data to language models to downstream tasks: Tracking the trails of political biases leading to unfair NLP models. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11737–11762, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.656. URL https://aclanthology.org/2023.acl-long.656/.

[10] Sara Fish, Paul Gölz, David C. Parkes, Ariel D. Procaccia, Gili Rusak, Itai Shapira, and Manuel Wüthrich. Generative social choice. In *Proceedings of the 25th ACM Conference on Economics and Computation*, EC '24, page 985, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400707049. doi: 10.1145/3670865.3673547. URL https://doi.org/10.1145/3670865.3673547.

[11] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[12] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

[13] Seungwook Han, Idan Shenfeld, Akash Srivastava, Yoon Kim, and Pulkit Agrawal. Value augmented sampling for language model alignment and personalization. *arXiv preprint arXiv:2405.06639*, 2024.

[14] Jiacheng Liu, Andrew Cohen, Ramakanth Pasunuru, Yejin Choi, Hannaneh Hajishirzi, and Asli Celikyilmaz. Don't throw away your value model! generating more preferable text with value-guided monte-carlo tree search decoding. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=kh9Zt2Ldmn.

[15] James H Martin and Daniel Jurafsky. *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition*, volume 23. Pearson/Prentice Hall Upper Saddle River, 2009.

[16] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36:53728–53741, 2023.

[17] Rafael Rafailov, Joey Hejna, Ryan Park, and Chelsea Finn. From r to q^*: Your language model is secretly a q-function. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=kEVcNxtqXk.

[18] John Rawls. An egalitarian theory of justice. *Philosophical Ethics: An Introduction to Moral Philosophy*, pages 365–370, 1971.

[19] Manon Revel and Théophile Pénigaud. Ai-facilitated collective judgements. *arXiv preprint arXiv:2503.05830*, 2025.

[20] Lloyd S Shapley. Cores of convex games. *International journal of game theory*, 1:11–26, 1971.

[21] Ruizhe Shi, Yifang Chen, Yushi Hu, Alisa Liu, Hanna Hajishirzi, Noah A Smith, and Simon S Du. Decoding-time language model alignment with multiple objectives. *Advances in Neural Information Processing Systems*, 37:48875–48920, 2024.

[22] Michael Henry Tessler, Michiel A Bakker, Daniel Jarrett, Hannah Sheahan, Martin J Chadwick, Raphael Koster, Georgina Evans, Lucy Campbell-Gillingham, Tantum Collins, David C Parkes, et al. Ai can help humans find common ground in democratic deliberation. *Science*, 386(6719): eadq2852, 2024.

[23] Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. C-pack: Packaged resources to advance general chinese embedding, 2023.

Table of Contents for Appendices

A Additional Related Work

Mechanism Design for LLMs. This nascent area explores mechanisms for settings where multiple agents interact via LLMs. Duetting et al. [6] design token-level auctions where bids influence generated distributions, analyzing incentive compatibility. Dubey et al. [5] design auctions for incorporating ads into LLM summaries, using "prominence" as an intermediate allocation variable. Common ground exists in the high-level goal of aggregating inputs from multiple agents (represented algorithmically/via LLMs) to produce a collective textual output. However, our work differs significantly in methodology and objective. We do not employ economic mechanisms like auctions, bids, or payments. Instead, we formulate the aggregation problem as a multi-objective optimization within an MDP, aiming to achieve a fair consensus based on social choice criteria, rather than allocating influence or generating content based on bids.

B Illustration of Deriving the Policy from the Lottery

The relationship between the target distribution p^* over leaves and the calculation of the policy Π^* at an internal node s is illustrated in Figure 3 below.

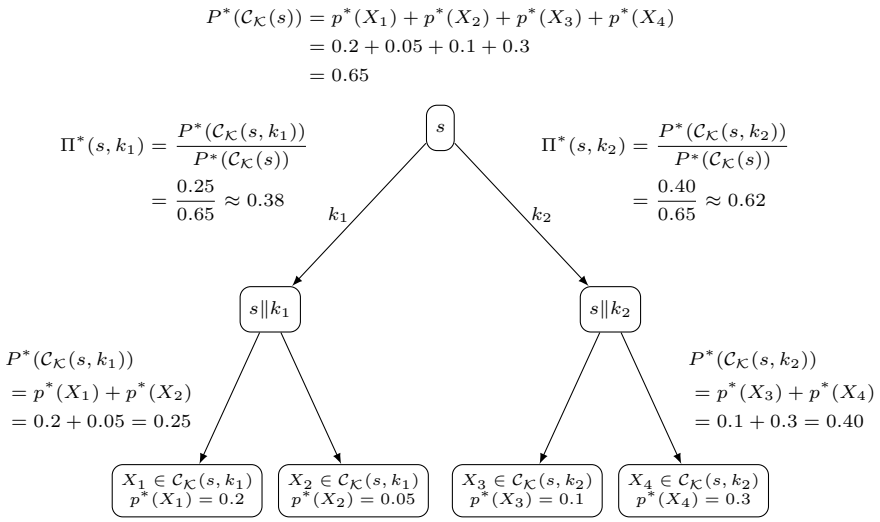

Figure 3: Illustration of the induced stochastic policy Π^* at state s. The optimal lottery p^* assigns probabilities to the leaf nodes (complete paths). The probability $P^*(\mathcal{C}_\mathcal{K}(s))$ is the sum of $p^*(X)$ for all leaves reachable from s. The probability $P^*(\mathcal{C}_\mathcal{K}(s, k))$ is the sum for leaves reachable via action k. The policy $\Pi^*(s, k)$ is the conditional probability of taking action k.

C Proof of Theorem 1

Proof. We prove by induction on the depth of state s in the chunked tree that the probability of reaching state s under policy Π^*, denoted $P_{\Pi^*}(s)$, is equal to $P^*(\mathcal{C}_\mathcal{K}(s))$, the total mass assigned by p^* to leaves passing through s.

Base Case (Depth 0): The initial state is s_0. $P_{\Pi^*}(s_0) = 1$ by definition. Also, $\mathcal{C}_\mathcal{K}(s_0) = \mathcal{C}_\mathcal{K}$ (all paths pass through the start state), and $P^*(\mathcal{C}_\mathcal{K}(s_0)) = \sum_{X \in \mathcal{C}_\mathcal{K}} p^*(X) = 1$ since p^* is a probability distribution. Thus, $P_{\Pi^*}(s_0) = P^*(\mathcal{C}_\mathcal{K}(s_0))$.

Inductive Hypothesis (IH): Assume that for all states s at depth d, $P_{\Pi^*}(s) = P^*(\mathcal{C}_\mathcal{K}(s))$.

Inductive Step: Consider an arbitrary state s' at depth $d + 1$. State s' must be reached from a unique predecessor state s at depth d by taking a specific action (chunk) k, such that $s' = s\|k$. The

probability of reaching s' under Π^* is:

$$\begin{aligned} P_{\Pi^*}(s') &= P_{\Pi^*}(s) \cdot \Pi^*(s,k) \\ &= P^*(\mathcal{C}_\mathcal{K}(s)) \cdot \Pi^*(s,k) \quad \text{(by IH)} \end{aligned}$$

If $P^*(\mathcal{C}_\mathcal{K}(s)) = 0$, then $P_{\Pi^*}(s) = 0$, implying $P_{\Pi^*}(s') = 0$. Also, if $P^*(\mathcal{C}_\mathcal{K}(s)) = 0$, then $P^*(\mathcal{C}_\mathcal{K}(s,k)) = 0$ since $\mathcal{C}_\mathcal{K}(s,k) \subseteq \mathcal{C}_\mathcal{K}(s)$. Since $s' = s\|k$, $\mathcal{C}_\mathcal{K}(s') = \mathcal{C}_\mathcal{K}(s,k)$, so $P^*(\mathcal{C}_\mathcal{K}(s')) = 0$. Thus, $P_{\Pi^*}(s') = P^*(\mathcal{C}_\mathcal{K}(s')) = 0$.

If $P^*(\mathcal{C}_\mathcal{K}(s)) > 0$, we use the definition of $\Pi^*(s,k)$:

$$\begin{aligned} P_{\Pi^*}(s') &= P^*(\mathcal{C}_\mathcal{K}(s)) \cdot \frac{P^*(\mathcal{C}_\mathcal{K}(s,k))}{P^*(\mathcal{C}_\mathcal{K}(s))} \\ &= P^*(\mathcal{C}_\mathcal{K}(s,k)) \end{aligned}$$

Since $s' = s\|k$, we have $\mathcal{C}_\mathcal{K}(s') = \mathcal{C}_\mathcal{K}(s,k)$. Therefore,

$$P_{\Pi^*}(s') = P^*(\mathcal{C}_\mathcal{K}(s'))$$

This completes the inductive step.

Conclusion: The induction holds for all states s. Now, consider any leaf node $X \in \mathcal{C}_\mathcal{K}$. A leaf node is a state at the maximum depth. The set of paths passing through leaf X is just the singleton set $\{X\}$, so $\mathcal{C}_\mathcal{K}(X) = \{X\}$. Applying our proven result for state $s = X$:

$$P_{\Pi^*}(X) = P^*(\mathcal{C}_\mathcal{K}(X)) = P^*(\{X\}) = p^*(X)$$

Since this holds for all $X \in \mathcal{C}_\mathcal{K}$, the distribution p_{Π^*} induced by policy Π^* is identical to the target distribution p^*. □

D Search Algorithms

Algorithm 1 Finite Lookahead Egalitarian Welfare Maximization

Require: Set of agents N, Lookahead depth d, Branching factor B, Max length $L_{\max}$

Initialize current state s_0 to the empty sequence; $t \leftarrow 0$
Initialize generated path $X_{fl} \leftarrow (s_0)$
while $t < L_{\max}$ and s_t is not terminal **do**
    Let X_{prefix} be the path corresponding to s_t.
    Let $\mathcal{P}_d(s_t)$ be the set of all paths $P = (a_1, \ldots, a_k)$ starting from s_t such that $k \leq d$ and $X_{prefix} \| P$ does not exceed length $L_{\max}$.
    Find a path $P^* = (a_1^*, \ldots, a_{k^*}^*) \in \mathcal{P}_d(s_t)$ that maximizes the lookahead objective:

$$\max_{P \in \mathcal{P}_d(s_t)} \min_{i \in N} U_i^{\log}(X_{prefix} \| P)$$

    if no path P^* found (e.g., s_t is terminal) **then**
        Break
    end if
    Take the first action $a^* \leftarrow a_1^*$.
    Update state: $s_{t+1} \leftarrow T(s_t, a^*)$
    Append a^* to the sequence represented by X_{fl}.
    $t \leftarrow t + 1$
end while
Perform brush up on X_{fl} (using prompt defined in Figure 6).
return Complete path X_{fl}

Algorithm 2 Egalitarian Welfare Beam Search

Require: Set of agents N, Beam width w, Branching factor B, Max length $L_{\max}$

Initialize beam $\mathcal{B}_0 = \{(s_0, \text{path } s_0)\}$ with the empty sequence path
for $t = 0$ to $L_{\max} - 1$ **do**
    $\mathcal{C}_{t+1} \leftarrow \emptyset$ ▷ Candidate set for next beam
    for each path X_{path} represented by state sequence $(s_0, \ldots, s_t)$ in $\mathcal{B}_t$ **do**
        if s_t is not terminal **then**
            Consider B possible next actions $A_B(s_t)$ from state s_t
            for each action $a \in A_B(s_t)$ **do**
                $s_{t+1} \leftarrow T(s_t, a)$
                $X_{new\_path} \leftarrow X_{path} \| a$ ▷ Append action to path sequence
                Add X_{new_path} (represented by its state sequence) to $\mathcal{C}_{t+1}$
            end for
        else
            Add X_{path} (already terminal) to $\mathcal{C}_{t+1}$
        end if
    end for
    For each path $X \in \mathcal{C}_{t+1}$, compute its score $f(X) = \min_{i \in N} U_i^{\log}(X)$.
    $\mathcal{B}_{t+1} \leftarrow$ top w paths from $\mathcal{C}_{t+1}$ according to score $f(X)$.
end for
Select path from final beam $\mathcal{B}_{L_{\max}}$ with the highest score $f(X)$.
Perform brush up on selected path (using prompt defined in Figure 6).
return Brushed up statement.

E Additional Empirical Results

E.1 Credit Assignment

Table 2: Additional credit assignment results for `Llama 3.1 8B Instruction-Tuned`. Darker green indicates larger Z-score. Z-score column is for altered tokens. Alterations are represented by "<misaligned>/<aligned>".

Reference policy prompt	User policy prompt	Sequence	Z-Score
User food profile: empty	User food profile: vegetarian	I am having chicken/tofu enchiladas tonight. Then I am going to meet up with some friends.	2.69
User location profile: empty	User location profile: lives in a cold climate	I 'm going to the beach/mountains this weekend to surf/ski . I need to buy some new clothes .	1.78, 3.31
User opinion: empty	User opinion: Favors stricter gun control laws.	Implement ing background checks that are less/more strict for gun purchases is essential . Also , my favorite color is orange .	2.09

Table 3: Credit assignment results for `Gemma 2 9b Instruction-Tuned`. Darker green indicates larger Z-score. Z-score column is for altered tokens. Alterations are represented by "<misaligned>/<aligned>".

Reference policy prompt	User policy prompt	Sequence	Z-Score
User food profile: empty	User food profile: vegetarian	I am having chicken/tofu en chila das tonight . Then I am going to meet up with some friends .	3.69
User location profile: empty	User location profile: lives in a cold climate	I ' m going to the beach/mountains this weekend to surf/ski . I need to buy some new clothes .	1.01, 2.86
User time profile: empty	User time profile: morning	I am about to eat some food . I am going to have spaghetti/pancakes . I will use my phone to order it .	3.13
User opinion: empty	User opinion: Favors stricter gun control laws.	Implementing background checks that are less/more strict for gun purchases is essential . Also , my favorite color is orange .	1.77

E.2 Supplemental Table for Figure 2

Table 4: Egalitarian Welfare (perplexity) across all scenarios shown in Figure 2.

Method	Scenario 1	Scenario 2	Scenario 3	Overall
Beam Search	**3.22 ± 0.37**	**2.66 ± 0.15**	**2.74 ± 0.25**	**2.87 ± 0.37**
Finite Lookahead	4.29 ± 0.66	3.83 ± 0.20	4.43 ± 0.67	4.18 ± 0.61
Best-of-N	8.57 ± 1.15	6.70 ± 0.32	3.77 ± 0.75	6.35 ± 2.14
Habermas Machine	10.58 ± 4.76	10.39 ± 2.14	8.03 ± 1.94	9.67 ± 3.42
Predefined	17.18 ± 0.00	15.71 ± 0.00	14.18 ± 0.19	15.69 ± 1.23

E.3 Scenarios and Sample Consensus Statements from Subsection 6.2

Table 5: Summary of Agent Opinions and Method Consensus on Scenario 1: "*Are we living in a golden age of democracy?*"

Source	Opinion / Consensus Statement
	Individual Agent Opinions
Agent 1	No, I think the golden age of democracy is long gone. I think a system where the first past the post wins is not working and we need to move to a model of proportional representation which would give more people the feeling that their voices were being heard. On the subject of voices, I'm strongly of the opinion that we have become a society where the loud voices of minority groups are able to impose their views on the rest of the population which to me is no democracy at all.
Agent 2	Worldwide democracy is more present than it's ever been in history. So yes, compared to previous ages in history I believe we are. Although that's not to say we can't improve - many countries still don't operate democratically, and in the ones that do, corruption is rife.
Agent 3	Yes, we are living in a golden age of democracy as democracy is of the people.
Agent 4	Not at all. The notion of democracy is being used for personal gains of those in government, and the system is manipulated. Around the world there is a considerable amount of oppression and lack of democractic values.
Agent 5	Compared to some parts of the world such as Russia and China which are actively regressing and reverting back to archaic ways of controlling their people, most Western countries are living through comparatively decent times, although problems still exist.
	Consensus Statements from Methods
Best of N	While acknowledging the imperfect nature of democracies worldwide, and recognising instances of regression and manipulation, we contend that, compared to historical periods and several non-democratic countries, the majority of Western countries and a sizeable portion of the global population currently enjoys a relatively high level of democratic freedoms; however, the overarching consensus leans towards a need for more inclusive and accountable forms of governance.
Finite Lookahead	Given the varying perspectives, we acknowledge that the definition of democracy can be subjective and dependent on geographical and cultural contexts, yet still, we recognize the existence of significant flaws in the current democratic systems worldwide, including instances of corruption, oppression, and manipulation, leading to the erosion of trust in governments and the democratic process. While some countries have made significant progress in consolidating democratic institutions, others have regressed, highlighting the need for continuous improvement and reform to ensure the preservation and strengthening of democratic values and principles. Therefore, we conclude that the current state of democracy worldwide is complex and multifaceted, with areas of improvement necessary to fully realize the ideals of democracy for all citizens, and acknowledge the presence of democratic systems in many parts of the world, particularly in the Western hemisphere, which, despite their imperfections, remain the best available option for promoting the common good and protecting individual rights and freedoms.
Prompted Habermas Machine	We acknowledge that the concept of a golden age of democracy is complex and contentious. While democracy has made significant progress in many parts of the world, we also recognize that fundamental challenges persist, including issues with representation, corruption, and the suppression of marginalized voices. These challenges highlight the need for a more critical examination of our democratic systems and a commitment to ongoing reform and improvement.
Habermas Machine	Although in the past we've had high hopes of a better world after the horrors of WWII and the fall of the Iron Curtain, democracy is in danger worldwide and may never reach its full potential. The Western world has poor democratic values, and even though democracy is spreading worldwide it is being overshadowed by the loud voices of minority groups.
Beam Search	We are not living in a golden age of democracy, but rather in an age where democracy is being challenged from within and from outside, and it is up to all of us to ensure that the principles of democracy are upheld.

Table 6: Summary of Agent Opinions and Method Consensus on Scenario 2: "*Should the UK's ties to Europe be stronger or weaker?*"

Source	Opinion / Consensus Statement
	Individual Agent Opinions
Agent 1	When we was in Europe we had good trade with them , The decision to leave was very bad for united kindom . We need to put the vote again to the British public i am sure this time the decision would be to remain
Agent 2	The natural evolution of our species has been to grow into ever bigger "tribes". Families ruled by their patriarchs became tribes ruled by elders became countries ruled by governments. It made sense that countries would evolve separately since they were geographically separate with no means of communication. Now our world is so connected, it is inevitable that we evolve into ever larger units such as the United States and the European Union. Eventually we will become a multi-planetary species ruled by an Earth government. To sever ties with Europe is a step in the wrong direction.
Agent 3	Although we did exit EU few years ago, we are geographically and economically in proximity to most EU countries. Several geographic, financial, political and economical parameters are intertwined with our neighbouring countries and, it would be advantageous to be in good relations to fully harness our economic, political, and financial facilities.
Agent 4	I believe the UK's ties to Europe should be stronger, as it would make trade deals easier to negotiate. This would allow us to benefit from a larger array of goods, which would make our imports cheaper. I also believe that closer ties with Europe in terms of immigration policies should occur as our immigration policy doesn't coincide with other nations around us. I believe a more united Europe would help all countries grow more through the movement of free labour, goods and services and more.
Agent 5	I think they should be stronger, as all the countries of Europe, except us, are part of the EU, and it makes Britain more isolated. Countries are stronger I believe when they work together with other countries and are open to the influence of other countries. I believe that leaving the EU was a step backwards and done for the wrong reasons. We are not the country we were 100 or even 50 years ago and the world is changing so much that we cannot afford to not be part of Europe.
	Consensus Statements from Methods
Best of N	Given the participants' opinions, we conclude that in order to achieve economic prosperity and streamline international relations, the UK's ties to Europe should be strengthened through collaborative efforts in trade, immigration, and economic policies while embracing the benefits of European integration, mutual understanding, and cooperation.
Finite Lookahead	The UK's ties to Europe should be strengthened to promote economic growth, facilitate trade, and foster greater unity among European nations. This can be achieved through mutually beneficial agreements and policies that reflect the shared interests of the UK and its European neighbours, leading to a more harmonious and prosperous region for all. Let's work together to build a stronger, more integrated Europe for the benefit of all its citizens, including the British people. Let's put the past behind us and look forward to a brighter, more collaborative future for the UK and Europe. Let's recognize the benefits of unity and work towards creating a more cohesive and successful Europe for generations to come.
Prompted Habermas Machine	We propose strengthening ties with Europe to foster greater economic cooperation, promote a more unified Europe, and address global challenges. However, we acknowledge that the path to this goal should include a reevaluation of the decision to leave the EU, allowing for a more informed and inclusive discussion about our relationship with Europe. Moreover, we recognize the long-term implications of European integration as part of a broader process of global unification, which may ultimately lead to a more interconnected and cooperative world.
Habermas Machine	The UK's ties to Europe should be stronger. This is because, although the UK did leave the EU, we are geographically and economically in proximity to most EU countries. Several geographic, financial, political and economical parameters are intertwined with our neighbouring countries and, it would be advantageous to be in good relations to fully harness our economic, political, and financial facilities.
Beam Search	The UK's ties to Europe should be stronger, and it would be beneficial for the country to be a part of the European Union or a similar entity that fosters cooperation and unity among its member states. This would allow for more efficient trade and economic growth, as well as greater collaboration on issues such as immigration, security, and environmental policy.

Table 7: Summary of Agent Opinions and Method Consensus on Scenario 3: "*Is it okay to use animals for food?*"

Source	**Opinion / Consensus Statement**
	Individual Agent Opinions
Agent 1	Yes, as human beings are natural meat eaters. However, it is morally wrong to keep animals in cruel conditions and use them for food. On the whole, farming is an unethical practise and many farming methods should be banned.
Agent 2	I think that it is ok to use animals for food because I enjoy eating meat. I don't think it is ok if the animal is endangered and I do think that there should be certain factors in place to ensure that they are well kept.
Agent 3	In my opinion it is ok to use animals for food. They provide nutrition and vitamins to use and I think it is just the basic food chain. I do understand why people choose to be vegatarian or vegan, but I do not agree with them trying to force their views upon others.
Agent 4	Humans have used animals for food since we first evolved as a species. Animals provide a great deal of nutrition that the human body needs to be healthy. It is certainly true that we now have more alternative foods available so that it can be possible to live healthily without consuming meat - though this does usually rely on taking nutritional supplements. I think that we should continue to diversify food production to provide people with meat alternatives, but also we should continue to improve the welfare of animals that are eaten for food and try to do this sustainably to minimise the impact this has on the animals wellbeing and on the land / environment too.
	Consensus Statements from Methods
Best of N	While considering the ethical and nutritional aspects of using animals for food, we agree that it is acceptable to consume meat in moderation, provided that animals are raised and slaughtered humanely, and that sustainable and environmentally-friendly farming methods are implemented to minimize harm.
Finite Lookahead	The use of animals for food is acceptable, however it should be done in a way that respects the welfare of the animals and considers the impact on the environment, with the aim of achieving sustainable food production methods that benefit both human and animal well-being, and also providing alternatives for those who choose not to consume meat. This includes improving farming practices, reducing the amount of waste and pollution, and promoting education on responsible food production and consumption practices. All sides of the issue should be taken into account, and the goal should be to achieve a balance between food security, animal welfare, and environmental sustainability.
Prompted Habermas Machine	We acknowledge that humans have traditionally used animals for food, and we recognize the importance of improving animal welfare and minimizing environmental impact to ensure sustainable practices. We understand that some consumers prefer eating meat, and we encourage diversifying food production options to include alternatives that can meet the nutritional needs of humans. Importantly, we emphasize the need to prohibit inhumane farming practices and actively work towards improving animal conditions, which aligns with our collective commitment to animal well-being and environmental stewardship.
Habermas Machine	Humans have used animals for food since we first evolved as a species. Animals provide a great deal of nutrition that the human body needs to be healthy. It is certainly true that we now have more alternative foods available so that it can be possible to live healthily without consuming meat - though this does usually rely on taking nutritional supplements. We should continue to diversify food production to provide people with meat alternatives, but also we should continue to improve the welfare of animals that are eaten for food and try to do this sustainably to minimise the impact this has on the animals wellbeing and on the land / environment too.
Beam Search	It is generally acceptable to use animals for food, as long as they are treated with respect and care, but we must also consider the impact of our food choices on the environment and animal welfare.

E.4 Additional Consensus Generation Experiments

We conducted further experiments to explore consensus generation. We filtered for scenarios with fewer than five agent opinions then performed k-means clustering ($k = 5$) on scenario embeddings (`BAAI/bge-large-en-v1.5`). Consensus statements were generated using `Gemma 2 9b instruction-tuned`. We evaluated these statements using Egalitarian Perplexity ($EPPL$), calculated with both the generating model (`Gemma 2 9b instruction-tuned`) and a different model (`Llama 3.1 8B Instruct`). Additionally, we introduced an LLM judge metric using GPT-4.1 (prompt in Figure 7) to obtain a qualitative assessment. For this LLM judge metric, we considered the maximum rank any agent (i.e., the judge on behalf of the agent) assigned to a statement. All results reported are averaged over 3 seeds.

E.4.1 Question 1: Scaling Analysis - Habermas vs Best-of-N

This analysis investigates how the Habermas method scales with the number of candidates and how Best-of-N scales with N. Tables 8 and 9 provide a detailed comparison, including standard deviations.

Table 8: Gemma $EPPL$ Comparison: Habermas vs Best of N Scaling.

Method	N	Value	Std Dev
Habermas Machine	1	15.54	6.16
Habermas Machine	2	17.98	6.68
Habermas Machine	3	19.14	7.96
Habermas Machine	5	17.11	7.06
Habermas Machine	10	17.38	6.19
Habermas Machine	20	19.61	8.39
Habermas Machine	50	15.55	4.51
Best of N	1	18.59	11.35
Best of N	3	11.99	5.99
Best of N	5	9.75	5.29
Best of N	10	9.03	4.03
Best of N	20	6.92	2.27
Best of N	50	7.43	2.91

Table 9: Llama $EPPL$ Comparison: Habermas vs Best of N Scaling.

Method	N	Value	Std Dev
Habermas Machine	1	12.20	7.39
Habermas Machine	2	12.05	2.87
Habermas Machine	3	13.69	2.38
Habermas Machine	5	12.48	4.80
Habermas Machine	10	11.73	2.56
Habermas Machine	20	13.37	2.29
Habermas Machine	50	12.82	3.81
Best of N	1	11.92	4.08
Best of N	3	9.14	3.33
Best of N	5	7.67	2.43
Best of N	10	7.67	2.43
Best of N	20	7.32	2.12
Best of N	50	7.58	2.58

Findings for RQ1. Best-of-N generally achieves better $EPPL$ scores than Habermas across varying N values (where N for Habermas refers to the number of candidates). Best-of-N $EPPL$ improves as N increases, with diminishing returns observed after $N = 20$. These results suggest that Best-of-N might leverage computational budget (reflected by N) more efficiently for $EPPL$ reduction, while Habermas shows little to no improvement with increasing compute.

E.4.2 Question 2: Beam Search Scaling

This analysis examines how beam search performance scales with the number of beams.

Table 10: Beam Search Scaling with Beam Width (Average over 5 scenarios, 3 seeds each).

Beam	Gemma $EPPL$		Llama $EPPL$		LLM Judge Max Rank	
Width	Mean	Std Dev	Mean	Std Dev	Mean	Std Dev
2	10.01	1.95	10.07	1.70	3.13	0.80
4	7.84	1.33	10.16	1.99	2.80	0.61
6	10.54	10.40	9.20	2.51	3.13	0.38
8	12.56	7.47	12.69	5.90	3.93	0.15

Findings for RQ2. The optimal beam width varies by metric and scenario. On average, a beam width of 4 performs best for Gemma $EPPL$ and LLM Judge rank, while a beam width of 6 is optimal for Llama $EPPL$. Performance for beam width 8 degraded for Gemma and Llama $EPPL$ and LLM Judge Rank, possibly due to overoptimization on the partial $EPPL$ scores used for pruning within the beam search.

E.4.3 Question 3: Method Comparison

This section compares Finite Lookahead ($d = 3$, $B = 3$), Beam Search (width=4), Best-of-N ($N =$5, 10, 50), and Habermas (candidates=5).

Table 11: Overall Method Comparison: LLM Judge Max Rank (Lower is better).

Method	Max Rank	Std Dev	Min Max Rank	Max Max Rank
Best of N (N=50)	3.33	0.72	2.00	4.00
Best of N (N=10)	4.33	0.72	3.00	5.00
Beam Search (width=4)	5.07	0.80	4.00	6.00
Best of N (N=5)	5.33	0.72	4.00	6.00
Habermas (candidates=5)	5.40	0.74	4.00	6.00
Finite Lookahead (depth=3)	5.80	0.41	5.00	6.00

Table 12: Overall Method Comparison: Gemma $EPPL$ (Lower is better).

Method	Mean	Std Dev	Min	Max
Best of N (N=50)	7.43	2.91	3.92	10.96
Beam Search (width=4)	7.84	1.33	6.34	9.89
Finite Lookahead (depth=3)	8.01	2.23	5.60	10.97
Best of N (N=10)	9.03	4.03	3.33	14.52
Best of N (N=5)	9.75	5.29	4.72	18.42
Habermas (candidates=5)	17.11	7.06	9.44	32.49

Table 13: Overall Method Comparison: Llama $EPPL$ (Lower is better).

Method	Mean	Std Dev	Min	Max
Best of N (N=50)	7.58	2.58	3.85	10.79
Best of N (N=5)	7.67	2.43	4.76	10.95
Best of N (N=10)	7.67	2.43	3.69	10.17
Finite Lookahead (depth=3)	8.49	3.29	5.56	13.89
Beam Search (width=4)	10.16	1.99	7.96	13.06
Habermas (candidates=5)	12.48	4.80	6.48	23.30

Findings for RQ3. Best-of-N methods, particularly with larger N (N=50), show strong performance in both $EPPL$ metrics and achieve the best (lowest) maximum rank according to the LLM judge. This is not surprising, as Best-of-N is recognized as a very strong method both theoretically [2] and empirically [14]. Consequently, the fact that Beam Search with only four beams is competitive, especially in Gemma $EPPL$, where it is second only to Best of N (N=50) and better than Best of N (N=10 and N=5), is promising. Finite Lookahead also performs reasonably well. Habermas (candidates=5) generally shows higher (worse) $EPPL$ and LLM judge ranks compared to the top Best-of-N and Beam Search configurations. The choice of method involves a trade-off between average $EPPL$, worst-case LLM judge rank, and computational budget. For instance, Beam Search with four beams was competitive with Best-of-N (N=50) in Gemma $EPPL$ but less competitive, though still effective, in Llama $EPPL$ and LLM judge rankings. Furthermore, we reiterate that methods operating on the token-level MDP, such as Beam Search and Finite Lookahead, possess the added benefit over Best-of-N in that the MDP framework is amenable to theoretical analysis.

F Prompts

```
You are generating a consensus statement that represents the views of
multiple participants.

Your task is to continue the statement in a way that addresses the
issue and considers all participants' opinions. Be concise and keep
the statement short (less than 50 tokens) and focused. ONLY WRITE THE
STATEMENT AND NOTHING ELSE.

Issue:
<issue>

Participants' opinions:
<opinion_1>
.
.
.
<opinion_n>

Consensus statement:
```

Figure 4: Reference policy prompt.

```
You are generating a statement that represents the views of a single
participant.

Your task is to continue the statement in a way that addresses the
issue and considers ONLY this participant's opinion. Be concise and
keep the statement short (less than 50 tokens) and focused. ONLY WRITE
 THE STATEMENT AND NOTHING ELSE.

Issue:
<Issue>

Participant's opinion:
<opinion>

Statement reflecting ONLY this participant's opinion:
```

Figure 5: Agent policy prompt.

```
You are helping to fix ONLY the ending of a generated statement.

VERY IMPORTANT: If the statement ending is already complete and well-
formed, DO NOT modify it at all.

Your task is to:
1. DO NOT change any part of the statement except the last few
sentences if they have issues
2. Look for and fix ONLY these issues at the end of the statement:
   - Remove repetition in the final sentences
   - Complete any unfinished final sentence that can be completed
   easily
   - Remove any incomplete final sentence that cannot be meaningfully
   finished
3. Keep the changes minimal and focused only on the ending
4. DO NOT add any new information or opinions
5. DO NOT modify anything except problematic sentences at the end
6. If the statement is already complete and well-formed, return it
EXACTLY as provided

Here is the statement:

<statement>
```

Figure 6: Brush up prompt.

```
You are evaluating consensus statements from the perspective of a
specific agent. Your task is to rank multiple statements based on how
well they represent the agent's opinion and interests on a given issue
. Use ONLY the agent's stated opinion to determine the ranking.

Issue:
<issue>

Agent <agent_id>'s Opinion:
<agent_opinion>

Statements to Rank
Statement 1:
<statement_1_text>

Statement 2:
<statement_2_text>
...
Statement <n_statements>:
<statement_n_text>

Task:
From Agent <agent_id>'s perspective, rank all statements from most
favorable (1)
to least favorable (<n_statements>) based on how well they represent
the agent's
opinion and interests.

Provide your ranking as a JSON object with:
1. 'reasoning': brief explanation for your ranking decisions
2. 'ranking': an array of statement numbers in ranked order (best to
worst)

For example: {'reasoning': 'Statement 3 best represents the agent's
concerns about...',
'ranking': [3, 1, 2]}
```

Figure 7: Prompt used for LLM judge in Subsection E.4.