
Social Choice Foundations for Simulation-Augmented Generation

Anonymous Authors¹

Abstract

Users increasingly turn to AI systems for normative assistance—guidance on what one ought to do or think—yet models are often opaque about whose viewpoints they represent. A promising approach is *simulation-augmented generation* (SAGE), which involves querying generative simulations of individuals in a target population at inference time, soliciting their open-ended judgments, and synthesizing them into a response while transparently reporting whose viewpoints are reflected. However, inference-time simulation raises acute *scalability constraints*. Since the key benefit of simulation is improved *representativeness*, the core challenge is scaling simulation without sacrificing representation. We introduce the first formalization of this problem, grounded in proportional clustering concepts from social choice theory. We prove that to represent a population of m humans, we need only create $n \ll m$ simulations of them, and need only dynamically query $k \ll n$ of those simulations at inference time, while still maintaining approximate proportional representation guarantees for the full population. We empirically validate that our inference-time algorithm yields better representation–efficiency trade-offs than baseline approaches.

1. Introduction

People increasingly turn to AI models for *normative assistance*—guidance on what one ought to do or think—such as seeking interpersonal advice, determining which political candidate to support, or evaluating academic work. Yet empirical work shows that models can guide users in ways that systematically diverge from how other humans would:

¹Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

exhibiting sycophancy in interpersonal advice (Cheng et al., 2026), systematically over-representing political parties beyond voter support for them (Miyazaki & Hall, 2026), and applying different evaluative criteria for academic work than human peer reviewers (Abdulhai et al., 2026). For normative questions—where there is no objective ground truth—models are opaque and often arbitrary about whose viewpoints they represent, leaving users unable to situate the guidance they receive within the broader opinion space.

In a concurrent position paper under double-blind review, Anonymous (2026b) argue that for normative questions, models should act as faithful mirrors of society: helping users access the range of perspectives they would hear from an appropriate reference population, while being transparent about whose perspectives are represented. To this end, they propose **simulation-augmented generation** (SAGE) (Anonymous, 2026b): at inference time, given a user prompt, the model queries generative simulations of individuals in a target population and synthesizes their answers into a final response. The user can see which population was queried and how representative the response is of that population.

Here we focus on a set of technical challenges for achieving the vision of SAGE. While much work in machine learning focuses on improving simulation fidelity (Wu et al., 2026; Naous et al., 2026), the shift to using simulations as an inference-time tool introduces an additional requirement: *scalability*. Specifically, user-facing models must satisfy three constraints: *latency*, responding to requests within acceptable timeframes; *compactness*, managing context windows to support extended multi-turn conversations; and *computational cost*, ensuring efficient resource utilization. Crucially, standard approaches to satisfying these constraints may interfere with representation. For example, naively summarizing agent outputs in order to compress context risks losing the key benefit of simulations: their ability to reflect diverse viewpoints in a desired target population. The central challenge, therefore, is **achieving scalability while maintaining representation**.

1.1. Overview of Approach

Our research question is this: given a population of m humans (e.g., millions), can we learn only n simulated agents (e.g., 1000), and then, for any given prompt, select just k of them (e.g., 10) to query while guaranteeing that their responses proportionately represent the viewpoints of the full population? If so, this directly addresses the scalability constraints faced by inference-time systems. Querying only k agents improves latency, reduces inference costs, and yields a compact set of responses that is feasible to fit in the context window. The challenge, then, is to identify *which* k agents to query—efficiently and with formal representation guarantees.

Formalizing representation. To formulate such guarantees, we turn to *social choice theory*—a field which has long studied collective decision-making under diverse preferences. However, while classical social choice theory largely focuses on discrete and finite sets of alternatives, in our setting agents choose among open-ended responses across a large and essentially unconstrained set of prompts. To formalize representation in this open-ended setting, we rely on embedding spaces. In particular, we assume access to an embedding space in which each agent prefers responses that are closer to their own. This type of *viewpoint embedding* space has been used in prior work on social choice for AI alignment (Ge et al., 2024) and deliberation (De et al., 2026), and methods for learning such spaces have recently been explored (Anonymous, 2026a).

With this geometric framing in hand, we can state the representation guarantee that we seek. Given a prompt p , we aim to select k agents to query so that their responses *proportionally represent* the viewpoints of the overall population. Specifically, when a large group of agents have similar ideal points, a proportionate number of the k selected responses should be close to these agents in the embedding space—ensuring that widely-held perspectives are reflected in the sampled responses.

Efficiently guaranteeing representation. We show that to represent a population of m humans, we only need to simulate a random sample of $n \ll m$ individuals where n grows logarithmically in m . To then reduce from needing to query n agents to querying just k agents at inference time, we devise an algorithm that first predicts approximate response embeddings of all n agents using a cheap model, applies proportional clustering algorithms to these predicted embeddings to select k representative agents, and then generates responses from those k agents only. This *select-then-generate* approach calls the expensive agent simulations only $k \ll n$ times, yet we prove that the resulting k responses satisfy approximate representation guarantees with respect to the full population.

1.2. Our Contributions

Below we state our technical contributions.

- Formal representation guarantees.** We provide the first formalization of representative simulation grounded in social choice theory, leveraging a *viewpoint embedding space* (Anonymous, 2026a) and defining representation through geometric properties via axioms from the computational social choice literature, chiefly *proportionally representative fairness*¹ (Aziz et al., 2024).
- Efficient proportional clustering.** In order to efficiently select k out of n viewpoint embeddings, we develop a faster implementation of the spatial expanding approval rules (SEAR)² algorithm, a proportional clustering algorithm which satisfies the representation axiom we study. Our implementation achieves an $O(n^2 \log n)$ amortized running time, improving upon the previous $O(k^2 n^4)$ bound, and making it practical for inference-time deployment.
- Efficient querying of simulations.** We show that selection and generation approximately commute: by cheaply predicting agents’ response embeddings and applying proportional clustering to these predictions, we can select k representative agents *before* generating responses. We prove that responses from these k agents inherit approximate representation guarantees with respect to the full population of n agents, enabling inference-time generation from $k \ll n$ agents while preserving representation.
- Scaling from simulations to populations.** Our ultimate target is proportional representation of an underlying human population of size m , which may be far larger than the simulated population size n . We show that it suffices to run our method on a simulated population with $n \ll m$ that simulates a random (possibly stratified) sample of the true population: specifically, n need grow only logarithmically in m , and when the embedding dimension d is small, n can depend only on d and be independent of m .
- Efficient evaluation of representation.** To compare our method against alternative selection procedures, one must be able to evaluate whether an arbitrary set of k responses satisfies the desired representation axioms. We show that exactly checking our representation axiom is computationally hard. To address this, we introduce a continuous *representation evaluation* measure

¹We also provide theoretical results guaranteeing the *no over-representation property* (NORP) (Jerrett & Anshelevich, 2025) in Appendix F.

²More precisely, we consider a variant of this algorithm, discussed in Section 3.2.

based on the approximation ratio of a stronger representation axiom,³ which we prove can be computed in time $O(n^2(\log n + k))$.

6. **Empirical validation.** We evaluate our approach on real data from a deliberative discussion platform, demonstrating that our inference-time algorithm satisfies representation guarantees at substantially higher rates than k -means and random selection baselines.

2. Related Work

Social choice. Starting with the work of Aziz et al. (2017), an influential line of work in social choice formalizes the notion of proportional representation in committee elections and participatory budgeting (Aziz et al., 2018; Brill & Peters, 2023). Generally speaking, the notions introduced therein ensure that every subset of agents that is sufficiently large, and whose preferences are sufficiently cohesive, would be represented in the outcome. Importantly, a similar approach to proportional representation has been explored in the context of clustering (Chen et al., 2019; Aziz et al., 2024; Kellerhals & Peters, 2024). The main difference from the work on proportionality in committee elections is that agent preferences are induced by distances in a metric space—a view that we also adopt. In particular, the proportionality notions that are key to this paper, PRF and mEJR+, build on the work of Aziz et al. (2024) and Kellerhals & Peters (2024), respectively.

Our work is similar in spirit to that of Fish et al. (2026) and Boehmer et al. (2025). Like us, their goal is to generate a slate of viewpoints that is formally representative of the views of a given set of agents. They do this through an algorithm that uses *discriminative queries* and *generative queries* as building blocks. However, their implementation of the latter type of queries, in particular, requires numerous calls to an LLM. In the context of SAGE, their approach gives an alternative method for realizing the representation criterion but blatantly fails the scalability criterion.

Pluralism in LLMs. SAGE (Anonymous, 2026b) aims to build models that serve as conduits for human perspectives. As such, it aligns with several adjacent threads in alignment that advocate for showing users a range of perspectives, including Overton pluralism in pluralistic alignment (Sorensen et al., 2024), reasonable pluralism (Fisher et al., 2025), and work on mitigating disempowerment (Sharma et al., 2026). While LLM-based simulations have been developed for other purposes—such as replicating social science experiments (Hewitt et al., 2024; Aher et al., 2023; Kolluri et al., 2025) or serving as user simulators for model evaluation (Yao et al., 2025)—SAGE

³It is a metric analogue of the discrete social choice axiom *extended justified representation plus* (EJR+) (Brill & Peters, 2023).

introduces unique scalability requirements by deploying simulations as inference-time calls in user-facing systems. A closely related precursor is *modular pluralism* by Feng et al. (2024), which queries k “community LLMs” at inference time to surface diverse perspectives. However, their approach relies on a fixed, pre-specified set of community models, which cannot adapt to the full diversity of user queries. Our work addresses this limitation: we maintain the scalability of querying only a small number of k agents while enabling adaptive, prompt-specific selection from a larger population of n simulated individuals, with formal guarantees that the selected agents proportionally represent the broader population.

3. Social Choice in the Viewpoint Embedding Space

In this section, we formalize how to select k agents whose outputs (proportionally) represent the population’s viewpoints.

Following prior work (Coombs, 1964; DeSarbo & Hoffman, 1987; Poole & Rosenthal, 1985; Anonymous, 2026a), we assume access to a *viewpoint embedding space*, where agents’ responses to a prompt are embedded into a shared latent space. Each agent’s embedding serves as their ideal point, and preferences are spatial: agents prefer viewpoints closer to their ideal point over those farther away.

3.1. Representation Guarantees

We now formally introduce the representation guarantee we use throughout the paper, adapting proportional clustering in computational social choice (Aziz et al., 2024; Chen et al., 2019). Our goal is to formalize when a slate of viewpoint embeddings proportionally represents groups of agents with similar viewpoints in viewpoint space.

Let \mathcal{N} be the set of agents with $|\mathcal{N}| = n$. Let \mathbb{S}^{d-1} be the unit sphere in \mathbb{R}^d . Fix agents’ viewpoint embeddings $\mathbf{v} = (v_1, \dots, v_n) \in (\mathbb{S}^{d-1})^n$, a target slate size $k \in \mathbb{N}$, where all embeddings are unit-normalized. Let Δ denote Euclidean distance in \mathbb{R}^d .

Let $X \subseteq \mathbb{S}^{d-1}$ be a slate of viewpoint embeddings with $|X| = k$. Since the slate has k slots for n agents, a group $S \subseteq \mathcal{N}$ of size $|S| \geq n/k$ has a proportional claim to one slot; more generally, if $|S| \geq \ell n/k$ —equivalently, if S comprises at least an ℓ/k fraction of the population—then S has a proportional claim to ℓ slots. We call such a group ℓ -large. Size alone, however, is not enough to determine representation. A large group may contain agents with widely different viewpoints, so its claim should depend also on how cohesive its viewpoint embeddings are. The more cohesive the group, the closer its representatives should be. Thus, for an ℓ -large group S , a slate item counts toward

the group’s claim only if it is sufficiently similar to the viewpoints in S .

To make this precise, we use the group’s worst-case internal similarity as its representation threshold. For a group $S \subseteq \mathcal{N}$, define $D(S) := \max_{i,j \in S} \Delta(v_i, v_j)$. Thus every pair of agents in S has distance at most $D(S)$, and smaller values of $D(S)$ indicate more cohesive groups. We say that S is *t-cohesive* if $D(S) \leq t$.

Proportionally Representative Fairness (PRF), introduced by Aziz et al. (Aziz et al., 2024), requires every sufficiently large and cohesive group to receive its proportional number of such representatives.

Definition 3.1 (Proportionally Representative Fairness). We say that a slate X satisfies *proportionally representative fairness* (PRF) if, for every ℓ -large and t -cohesive group of agents $S \subseteq \mathcal{N}$ with $\ell \in [k]$, it holds that

$$\left| \left\{ x \in X \mid \min_{i \in S} \Delta(x, v_i) \leq t \right\} \right| \geq \ell.$$

3.2. Efficient and Proportional Clustering

To compute a slate X consisting of k representative embeddings satisfying our representation guarantees, we adapt a *proportional* clustering algorithm—a variant of *spatial expanding approval rule* (SEAR)⁴—to our setting which, unlike (balanced) k -means, provably returns a set of k centroids satisfying PRF.

SEAR. The input to SEAR is the collection of n agents’ response embeddings v_1, \dots, v_n and a nonnegative integer k . Each of these initially has a budget of $b_i = 1$ for $i \in \mathcal{N}$. The slate is initialized to $X = \emptyset$. The algorithm grows Euclidean balls around viewpoint embeddings at uniform speed. As soon as a ball around some agent i ’s embedding v_i covers points S whose total budget is at least n/k , then the budget of agents in S is reduced by n/k in total⁵ and $X = X \cup \{v_i\}$. The algorithm terminates and returns the slate X once it reaches size k .

Runtime. From the algorithmic perspective, the main issue is efficiency, since we ultimately want to enforce representation guarantees at inference time while respecting latency constraints. Aziz et al. (2024) show that a closely re-

⁴The proportional clustering literature contains several closely related variants of this procedure of greedily picking groups to be represented and deactivating them, sometimes referred to as *Greedy Capture*. Although these variants share the same basic structure, the precise implementation matters for the axiomatic guarantees obtained. We discuss these algorithmic differences in the appendix.

⁵In some way, this does not matter for PRF, but has implications for runtime.

lated precursor of this algorithm⁶ has running time $O(k^2 n^4)$. For our application, this would be prohibitive even for simulated populations of moderate size, such as $n \approx 10^3$.

We show that SEAR admits a substantially faster and more streamlined implementation, which is crucial for our inference-time setting and makes proportional clustering practical in our framework.

Theorem 3.2. *SEAR can be implemented with amortized running time $O(n^2 \log n)$.*

In particular, we distribute the payment so that a minimal set $S' \subseteq S$ pays n/k budget: All except perhaps one $v \in S$ furthest away from v_i spend their entire budget, while v ’s budget is reduced to $b_v = n/k - \lfloor n/k \rfloor$. By additionally only growing balls around the responses of agents whose remaining budget is positive, the algorithm avoids overrepresenting agent viewpoints in the slate. We discuss this in more detail in Appendix F.

4. Selection and Generation Commute

Taken together, the ideas from the previous section give a conceptually complete inference-time procedure for open-ended viewpoint representation. Given a vocabulary set \mathcal{V}^* , for each agent $i \in \mathcal{N}$, let $A_i : \mathcal{V}^* \rightarrow \mathcal{V}^*$ denote the agent simulation: given a prompt p , it generates agent i ’s viewpoint. We write $o_i^* := A_i(p)$ when p is clear from context. Let $\psi : \mathcal{V}^* \rightarrow \mathbb{S}^{d-1}$ be an embedding function, which maps responses to the viewpoint embedding space. This immediately yields a baseline procedure, which we call *generate-then-select*.

Generate-then-select In the *generate-then-select* paradigm, given a prompt p , we generate every agent’s response $o_i^* = A_i(p)$, embed the generated responses as $\{\psi(o_i^*)\}_{i \in \mathcal{N}}$, run SEAR in this embedding space to select k representative agents $R \subseteq \mathcal{N}$, and then return the corresponding response slate $Y := \{o_i^* : i \in R\}$.

This pipeline establishes conceptual feasibility, but its main limitation is computational: it requires one generation per agent. If the simulations A_i are implemented as large autoregressive models, especially ones using inference-time reasoning, then applying *generate-then-select* to large simulated populations becomes prohibitively expensive in both costs and latency. The goal is therefore to retain the same representational benefits with greater scalability.

Select-then-generate. The key idea is to replace explicit simulation with a cheaper embedding prediction step. Rather than generating a response from every agent and

⁶More precisely, the implementation analyzed in Aziz & Lee (2020) satisfies PRF but not NORP.

then embedding the resulting viewpoints, we learn a prompt-side embedding function $\phi_i(p)$ that directly predicts the embedding of the viewpoint that agent i would generate in response to p . We posit that predicting the viewpoint embeddings are easier to predict than predicting the full textual response, and cheaper in the sense that, unlike running the full simulation, the predictive model can be smaller and does not involve autoregressive generation or inference-time reasoning.

Definition 4.1 (γ - Proxy Viewpoint Embeddings). $\{\phi_i(p)\}_{i \in \mathcal{N}}$ are γ -proxy viewpoint embeddings if and only if there exists $\gamma \leq 2$ such that for any agent $i \in \mathcal{N}$, $\Delta(\phi_i(p), \psi(o_i^*)) \leq \gamma$.

Thus, proxy viewpoint embeddings approximate the ideal viewpoint geometry in a pointwise sense: each predicted embedding $\phi_i(p)$ is uniformly aligned with its corresponding ideal embedding $\psi(o_i^*)$, with distance at most γ .

This enables us to move representative selection before simulation, and replace the *generate-then-select* pipeline with a *select-then-generate* pipeline. Instead of first simulating all agents, we first compute the predicted viewpoint embeddings $\{\phi_i(p)\}_{i \in \mathcal{N}}$, run SEAR in this latent space, and query only the selected simulation models.

Select-Then-Generate

Input: prompt p , agents \mathcal{N} , predicted embedding maps $\{\phi_i\}_{i \in \mathcal{N}}$, simulation models $\{A_i\}_{i \in \mathcal{N}}$, slate size k .

Output: slate Y of k generated viewpoints.

1. **Predict embeddings.** Compute $\{\phi_i(p)\}_{i \in \mathcal{N}}$.
2. **Select representatives.** Run SEAR on $\{\phi_i(p)\}_{i \in \mathcal{N}}$ to obtain a set $X \subseteq \{\phi_i(p)\}_{i \in \mathcal{N}}$ of representative embeddings. Let $R \subseteq \mathcal{N}$ be the set of k corresponding agents.
3. **Simulate k selected agents.** Generate $A_i(p)$ only for agents $i \in R$.
4. **Return slate.** Return $Y := \{A_i(p) : i \in R\}$.

With a uniform upper bound of γ , *select-then-generate* can perform selection in the predicted embedding space before running any expensive simulations. The following theorem shows that this pipeline, which uses only k simulation queries, preserves the desired representation guarantee up to the approximation controlled by γ and t .

Theorem 4.2 (Amortized Selection–Generation Commuta-

tion). Let $\{\phi_i(p)\}_{i \in \mathcal{N}}$ be γ -proxy viewpoint embeddings. For every ℓ -large and t -cohesive group of agents $S \subseteq \mathcal{N}$ there exist at least ℓ outputs $o \in Y$ such that for each such o , there exists some agent $i \in S$ satisfying

$$\Delta(\psi(o), \psi(o_i^*)) \leq t + 2\gamma.$$

Thus, the slate Y of responses returned by the select-then-generate algorithm retains approximate representation guarantees with respect to the true viewpoint embeddings $\{\psi(o_i^*)\}_{i \in \mathcal{N}}$ while using only k simulations. Theorem 4.2 formalizes the sense in which selection and generation approximately commute: representative selection can be performed in the predicted embedding space before any simulation queries are made. In Appendix F.3, we describe how the viewpoint embedding space and the predicted embeddings could be jointly learned via *personalized reward models* and establish the resulting reward guarantees .

Population guarantees from sampled simulations.

Select-then-generate reduces the number of expensive simulation queries from n to k by selecting representatives before generation. However, the method still requires a pool of simulated agents whose predicted embeddings $\{\phi_i(p)\}_{i \in \mathcal{N}}$ are used for representative selection. Taking this pool to contain one simulation for every member of the population we wish to represent may be infeasible. For example, representing the entire U.S. population would require maintaining simulations for roughly 340 million people, which is impractical both computationally and operationally.

Fortunately, it is not necessary to instantiate a simulation for every person in the population: for representation guarantees, it suffices to simulate only a small, random sample of them. The following theorem shows that if this sampled set of simulations satisfies PRF, then the guarantee transfers approximately to the full human population. This result is algorithm-agnostic: it applies to any slate satisfying PRF on the sampled simulations.

Theorem 4.3 (Population Guarantees from Sampled Agents). Let \mathcal{M} denote the underlying population, with $|\mathcal{M}| = m$. Let $\mathcal{N} \subseteq \mathcal{M}$ be a sample of n agents drawn uniformly at random without replacement. If $n = \Omega(\frac{1}{\epsilon^2} \log \frac{km}{\delta})$ and Y^* is a PRF slate with respect to the sampled agents \mathcal{N} , then with probability at least $1 - \delta$, for every $\ell \in [k]$ and every t -cohesive group $S \subseteq \mathcal{M}$ with $|S| \geq (\ell/k + \epsilon)m$,

$$|\{o \in Y^* \mid \exists i \in S \text{ s.t. } \Delta(\psi(o_i^*), \psi(o)) \leq 2t\}| \geq \ell.$$

Thus, to represent a population of size m , it suffices to run select-then-generate on a much smaller sample of $n \ll m$ simulations (and then generate viewpoints only for the k selected representatives). In particular, the required sample

size grows only logarithmically in m . In Appendix E, we show that when d (the dimension of the embedding space) is small, n can depend only on d and be independent of m and k .

Note that Theorem 4.3 applies to arbitrary subsets $S \subseteq \mathcal{N}$, which requires approximating such groups by sampled proxy sets and leads to a geometric loss. Stronger guarantees are possible for *structured* cohesive groups, such as those induced by balls, boxes, or other low-complexity regions in the embedding space. When these regions form a bounded-VC range family, uniform convergence ensures that the sample preserves the size of every such group. Thus, PRF transfers from the sample to the population with *no additional loss* in the cohesiveness parameter t . Details can be found in Appendix E.3.

5. Experiments

Overview. Our goal is to enable a critical component of simulation-augmented generation: routing queries to agents in a way that is both *scalable* and *representative*. Thus far, we have developed the necessary theoretical components. In Section 4, we proved that generation and selection can be commuted, which allows us to query only $k \ll n$ agents while still maintaining approximate proportional representation guarantees. In particular, any algorithm that satisfies PRF under *generate-then-select* also yields approximate PRF under *select-then-generate*. We now present an end-to-end empirical demonstration of this commutation. Starting from a set of simulated agents, we train a viewpoint-embedding predictor and compare selection algorithms under both the *generate-then-select* and *select-then-generate* paradigms.

We evaluate the following three selection algorithms: SEAR, because it satisfies PRF; standard k -means, because it is a ubiquitous clustering baseline in practice; and random selection, because it provides simple distributional guarantees (e.g., each agent is selected with uniform probability). Under *select-then-generate*, we find that SEAR consistently outperforms both k -means and random selection in terms of proportional representation. Interestingly, although k -means does not theoretically guarantee PRF, for small k , it performs comparably to SEAR under *generate-then-select*; but when we reverse the order, SEAR substantially outperforms k -means, suggesting that the theoretical gap becomes more consequential when efficiency is a constraint and approximate response embeddings must be used.

Simulations. In simulation-augmented generation, we seek to query simulations for *open-ended* viewpoints. We, thus, base our simulations on data from Remesh (Konya et al., 2023; 2025), a collective response platform where participants share their open-ended viewpoints on questions

and vote on whether they agree or disagree with other participants’ viewpoints. Specifically, we use data from an open-source Remesh session in which $n = 310$ representative⁷ U.S. participants shared their viewpoints on campus protests related to the Israel/Palestine conflict.⁸ We construct one simulated agent per participant. Each agent is implemented using Llama-3.1-8B-Instruct (Grattafiori et al., 2024) and prompted to roleplay as the corresponding participant: the prompt includes the participant’s demographic profile⁹, along with their written viewpoints and binary agree/disagree votes on other participants’ opinions as few-shot context (see Appendix G.5 for the full prompt).

We do not purport that our simulated agents are faithful representations of the actual viewpoints of the underlying participants. Rather, they serve as plausible synthetic personas, following the common practice of creating simulations by prompting them with demographic and attitudinal profiles (Hewitt et al., 2024; Cui et al., 2025; Argyle et al., 2023). The goal of these experiments is not to create or evaluate the fidelity of individual-level simulation—a challenging problem studied extensively elsewhere (Hu et al., 2026; Zhou et al., 2026; Dou et al., 2025)—but rather to examine how different agent selection methods affect representation of the overall population. In this sense, our setup provides a controlled testbed for studying the scalability-representation tradeoff central to simulation-augmented generation.

Prompts. For evaluation, we generate 1,000 new prompts asking for an open-ended viewpoint. Specifically, we first prompt Claude Opus 4.6 to generate ten topics that could be predictable from participants’ responses to the original Remesh questions, and then to generate 100 diverse open-ended opinion questions per topic, yielding 1,000 prompts in total. Finally, we split the prompts into 50/10/40 train/validation/test sets. Appendix G.2 provides the full generation prompts, the resulting topics, and example questions for each topic.

Embedding prediction. Our representation guarantees assume an underlying *viewpoint embedding space* in which, for a given prompt, each agent has an ideal response and prefers responses that are closer in distance to their own. In this proof of concept, we use the off-the-shelf Qwen3-Embedding-0.6B (Zhang et al., 2025) model as the ground-truth embedder; future work could explore additional fine-tuning to further align the embedding distances with prefer-

⁷Participants were recruited via stratified sampling to be demographically representative of the general U.S. population. While not the focus of these experiments, our stratified-sampling results in Appendix E also bound how representation guarantees extend beyond the n participants to the U.S. population more broadly.

⁸The data are available at <https://github.com/akonya/polarized-issues-data> under a CC-BY-4.0 license.

⁹Gender, age, political affiliation, religious affiliation, education, urbanicity, and household income.

ence judgments (Anonymous, 2026a). To create the embedding predictor needed for *select-then-generate*, we fine-tune a Qwen3-Embedding-0.6B model to predict the ground-truth response embeddings from a description of the agent and the prompt.¹⁰ The description of each agent is generated by Claude 4.6 Opus summarizing the agent’s responses on the training prompts into a per-topic synopsis (see Appendix G.6 for the exact prompt). For training, we use a loss with two terms: (i) an MSE loss that encourages the predicted embedding to be close to the true embedding, and (ii) a Pearson correlation loss that encourages the predictor to preserve the pairwise distance structure among agent’s embeddings. Training details are provided in Appendix G.7.

Evaluating representation. Under both the *select-then-generate* and *generate-then-select* paradigms, the final output returned to the model is a slate of $k \ll n$ agent responses. Our goal is to measure how well these k responses represent all n agents. Unfortunately, we prove in Theorem B.1 that checking whether a given slate of responses satisfies PRF is NP-hard, even in a structured metric setting such as ours. For that reason, we turn to a *stronger* axiom—metric Extended Justified Representation+ (mEJR+)—defined formally in Definition C.1.¹¹ While mEJR+ implies PRF (Proposition C.2), it can be validated efficiently: we prove that Algorithm 2 in Appendix C runs in $O(n^2(\log n + k))$ time and provably computes how well the input slate of responses approximates mEJR+ for the agent population. The output is represented by an approximation parameter¹² $\beta > 0$ in which $\beta \leq 1$ implies the axiom is exactly satisfied, and a larger value of β implies that the slate is farther from

¹⁰The model takes in both the agent’s description and the current prompt, formatted as {agent_description} Question: {prompt}. Because the agent description is fixed across prompts, it can be treated as an upfront “prefill”: we cache the KV state for the user tokens once, and for each new prompt only process the prompt tokens.

¹¹To the best of our knowledge, mEJR+ has not been explicitly studied in the prior literature. However, it can be obtained by plugging the EJ+ axiom of Brill & Peters (2023, Def. 10) for approval-based elections into the general framework introduced by Kellerhals & Peters (2024, Def. 4) for adapting approval-based proportionality axioms to the metric setting.

¹²Recall our intuition from Section 3.1 that a set of agents S of size at least $\ell n/k$ should have some “claim” over ℓ responses in the slate, where the strength of the claim depends on how cohesive S is. PRF posits that if each agent in S is at most a distance of D away from every other agent in S , then each of these ℓ responses should be at most D away from some agent in S (Def. 3.1). mEJR+, on the other hand, requires that if agents in S are all at most D away from a shared agent, these ℓ responses should all be at most D away from a specific agent in S . In other words, mEJR+ has a stronger representation requirement for a softer cohesiveness condition. Accordingly, the approximation parameter β for mEJR+ is the smallest number such that for every such set S , there is some agent in S that is βD away from their ℓ th favorite response in the slate.

the mEJR+ guarantees.

Results. We evaluate three selection methods—SEAR, k -means¹³ and random selection—under the *generate-then-select* and *select-then-generate* paradigms. For each prompt in the test set, and $k \in \{3, \dots, 10\}$, we compute the mEJR+ approximation ratio β of the resulting slate of k responses, which measures how well the slate represents all n agents’ viewpoints on that prompt (Figure 1).

Under the *generate-then-select* paradigm, SEAR nearly always satisfies mEJR+ in practice, despite lacking a theoretical guarantee—since SEAR only satisfies PRF, and mEJR+ is a strictly stronger condition that implies PRF. For small values of $k \leq 5$, k -means also satisfies mEJR+ on $\geq 98\%$ of prompts. However, its performance degrades as k increases: by $k = 10$, k -means satisfies mEJR+ for only 79% of prompts, compared to 99% for SEAR.

Under *select-then-generate*, both SEAR and k -means perform worse than in *generate-then-select*, as expected when using approximate embeddings. However, the performance of k -means degrades far more than SEAR, often falling below random selection. Across all values of k , SEAR satisfies mEJR+ on 11-29 percentage points more prompts than k -means and on 4-23 percentage points more prompts than random selection. These results indicate that SEAR is more robust to approximate embeddings than k -means, suggesting that theoretically grounded selection methods are essential for maintaining representativeness in efficiency-constrained settings.

6. Discussion

In this work, we provided the first formalization of scalable and representative routing for simulation-augmented generation, grounded in social choice theory. We showed that to represent a population of m humans, we need only create $n \ll m$ simulations of them, and need only dynamically query $k \ll n$ of those simulations at inference time, while still maintaining approximate proportional representation guarantees for the full population.

Limitations. Several limitations warrant discussion. Our empirical evaluation uses only one dataset, leaving generalization to other domains and populations as future work. In our proof-of-concept, we employ off-the-shelf embedding models rather than learning preference embeddings tailored to the task (Anonymous, 2026a). Furthermore, although we prove that selection and generation can be approximately commuted, representation guarantees degrade as k grows.

¹³We run standard k -means on the response embeddings: ground-truth embeddings for *generate-then-select* and predicted embeddings for *select-then-generate*. We then select the k agents whose embeddings are closest to the k centroids.

Effect of Agent Selection Algorithm on Slate Representativeness

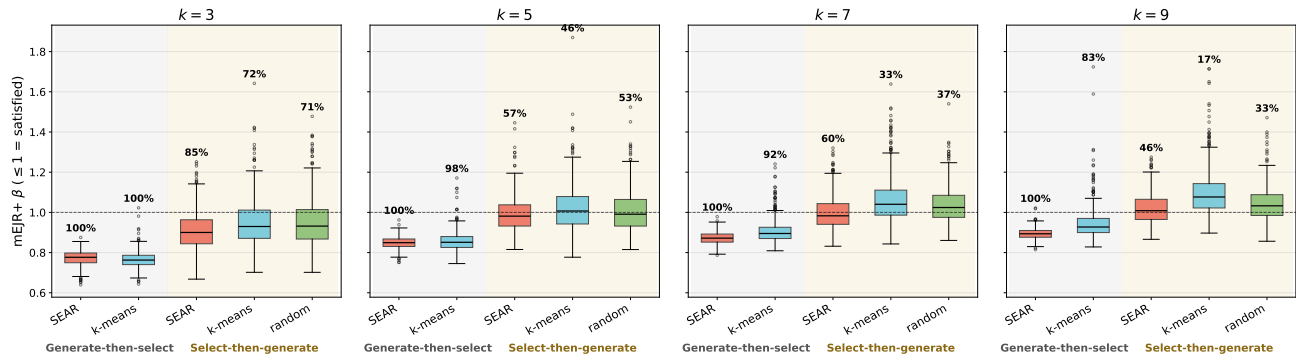


Figure 1. Distribution of the mEJR+ β approximation ratio across test prompts for each selection method and slate size $k \in \{3, \dots, 10\}$ (results for all k can be found in Figure G.1). A ratio ≤ 1 (dashed line) indicates that exact mEJR+ is satisfied; percentages above each boxplot denote the proportion of prompts satisfying exact mEJR+. Since random selection produces identical results under *generate-then-select* and *select-then-generate*, we group it under *select-then-generate*. Appendix G.8 shows examples of the slates generated by each method.

Finally, as is standard in work on committee elections, the choice of k is left open and must be tailored to the application context, since it directly determines the threshold group size required for representation.

Future work. Several directions remain open. First, future work could further push the Pareto frontier between scalability and representativeness, including exploring methods that guarantee stronger axioms than PRF. Second, we did not study simulation fidelity. In practice, simulations will vary in quality, and different simulated individuals may be more reliable in different domains (e.g., politics versus interpersonal advice). How should routing account for heterogeneous fidelity? Finally, while this work focused on the problem of routing queries to agents, there remain several other research directions required for achieving the full vision of SAGE outlined by Anonymous (2026b), including developing higher-fidelity simulations, doing high-quality synthesis on top of simulation outputs, and interpretably communicating representation guarantees to users.

Broader impacts. Our work is situated within the broader context of simulation-augmented generation (SAGE), which aims to restore agency to users on normative questions by being transparent about whose viewpoints a model represents. That said, the specific contribution of this paper addresses only one component of this vision: the routing problem of efficiently selecting representative agents at inference time. A trustworthy implementation of SAGE requires several additional components—including high-fidelity simulation, confidence estimation, and appropriate user interfaces. We caution against deploying the routing methods developed here in isolation, without adequate attention to these complementary elements.

References

- Abdulhai, M., White, I., Wan, Y., Qureshi, I., Leibo, J., Kleiman-Weiner, M., and Jaques, N. How LLMs Distort Our Written Language. *arXiv preprint arXiv:2603.18161*, 2026.
- Aher, G., Arriaga, R. I., and Kalai, A. T. Using Large Language Models to Simulate Multiple Humans and Replicate Human Subject Studies. In *Proceedings of the 40th International Conference on Machine Learning, ICML’23*. JMLR.org, 2023.
- Anonymous. Embeddings for Preferences, Not Semantics. 2026a.
- Anonymous. For Questions of Ought, AI Could Use Some SAGE Advice. 2026b.
- Anthropic. Claude opus 4.6 system card. Technical report, Anthropic, February 2026. URL <https://www-cdn.anthropic.com/0dd865075ad3132672ee0ab40b05a53f14cf5288.pdf>. Accessed: 2026-05-05.
- Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J. R., Rytting, C., and Wingate, D. Out of One, Many: Using Language Models to Simulate Human Samples. *Political Analysis*, 31(3):337–351, 2023.
- Aziz, H. and Lee, B. E. The expanding approvals rule: improving proportional representation and monotonicity. *Social Choice and Welfare*, 54(1):1–45, 2020.
- Aziz, H., Brill, M., Conitzer, V., Elkind, E., Freeman, R., and Walsh, T. Justified representation in approval-based committee voting. *Social Choice and Welfare*, 48(2): 461–485, 2017.

- 440 Aziz, H., Elkind, E., Huang, S., Lackner, M., Sánchez-
441 Fernández, L., and Skowron, P. On the complexity of
442 extended and proportional justified representation. In
443 *Proceedings of the 32nd AAAI Conference on Artificial*
444 *Intelligence*, pp. 902–909, 2018.
- 445 Aziz, H., Lee, B. E., Chu, S. M., and Vollen, J. Proportion-
446 ally representative clustering. In *Proceedings of the 20th*
447 *Conference on Web and Internet Economics*, 2024.
- 449 Bennett, J. F. and Hays, W. L. Multidimensional unfolding:
450 Determining the dimensionality of ranked preference data.
451 *Psychometrika*, 25(1):27–43, 1960.
- 453 Black, D. On the rationale of group decision-making. *Jour-*
454 *nal of Political Economy*, 56(1):23–34, 1948.
- 455 Boehmer, N., Fish, S., and Procaccia, A. D. Generative
456 social choice: The next generation. In *Proceedings of*
457 *the 42nd International Conference on Machine Learn-*
458 *ing*, 2025. URL [https://openreview.net/forum?](https://openreview.net/forum?id=E1E6T7KH1R)
459 [id=E1E6T7KH1R](https://openreview.net/forum?id=E1E6T7KH1R).
- 461 Bogomolnaia, A. and Laslier, J.-F. Euclidean preferences.
462 *Journal of Mathematical Economics*, 43(2):87–98, 2007.
463 doi: 10.1016/j.jmateco.2006.09.004.
- 465 Brill, M. and Peters, J. Robust and verifiable proportionality
466 axioms for multiwinner voting. In *Proceedings of the*
467 *ACM Conference on Economics and Computation (EC)*,
468 2023.
- 469 Chen, X., Fain, B., Lyu, L., and Munagala, K. Proportionally
470 fair clustering. In *Proceedings of the 36th International*
471 *Conference on Machine Learning (ICML)*, 2019.
- 473 Cheng, M., Lee, C., Khadpe, P., Yu, S., Han, D.,
474 and Jurafsky, D. Sycophantic AI decreases prosocial
475 intentions and promotes dependence. *Science*,
476 391(6792):eaec8352, 2026. doi: 10.1126/science.
477 aec8352. URL [https://www.science.org/doi/](https://www.science.org/doi/abs/10.1126/science.aec8352)
478 [abs/10.1126/science.aec8352](https://www.science.org/doi/abs/10.1126/science.aec8352).
- 479
- 480 Coombs, C. H. *A Theory of Data*. Wiley, New York, 1964.
- 481
- 482 Cui, Z., Li, N., and Zhou, H. A large-scale replica-
483 tion of scenario-based experiments in psychology and
484 management using large language models. *Nature*
485 *Computational Science*, 5(8):627–634, 2025. doi: 10.
486 1038/s43588-025-00840-7. URL [https://doi.org/](https://doi.org/10.1038/s43588-025-00840-7)
487 [10.1038/s43588-025-00840-7](https://doi.org/10.1038/s43588-025-00840-7).
- 488
- 489 De, S., Gelauff, L., Goel, A., Milli, S., Procaccia, A. D.,
490 and Siu, A. Question the questions: Auditing repre-
491 sentation in online deliberative processes. In *Proceed-*
492 *ings of the ACM Web Conference 2026*, WWW ’26, pp.
493 1640–1650, New York, NY, USA, 2026. Association for
494 Computing Machinery. ISBN 9798400723070. doi:
10.1145/3774904.3792474. URL [https://doi.org/](https://doi.org/10.1145/3774904.3792474)
10.1145/3774904.3792474.
- DeSarbo, W. S. and Hoffman, D. L. Constructing MDS
joint spaces from binary choice data: A multidimensional
unfolding threshold model for marketing research. *Jour-*
nal of Marketing Research, 24(1):40–54, February 1987.
doi: 10.1177/002224378702400104.
- Dou, Y., Galley, M., Peng, B., Kedzie, C., Cai, W., Ritter, A.,
Quirk, C., Xu, W., and Gao, J. SimulatorArena: Are User
Simulators Reliable Proxies for Multi-Turn Evaluation of
AI Assistants?, 2025. URL [https://arxiv.org/abs/](https://arxiv.org/abs/2510.05444)
2510.05444.
- Elkind, E., Lackner, M., and Peters, D. Preference restric-
tions in computational social choice: A survey. *arXiv*
preprint arXiv:2205.09092, 2022.
- Enelow, J. M. and Hinich, M. J. *The spatial theory of voting:*
An introduction. CUP Archive, 1984.
- Feng, S., Sorensen, T., Liu, Y., Fisher, J., Park, C. Y.,
Choi, Y., and Tsvetkov, Y. Modular Pluralism: Plural-
istic Alignment via Multi-LLM Collaboration. In Al-
Onaizan, Y., Bansal, M., and Chen, Y.-N. (eds.), *Pro-*
ceedings of the 2024 Conference on Empirical Meth-
ods in Natural Language Processing, pp. 4151–4171,
Miami, Florida, USA, November 2024. Association
for Computational Linguistics. doi: 10.18653/v1/2024.
emnlp-main.240. URL [https://aclanthology.org/](https://aclanthology.org/2024.emnlp-main.240/)
2024.emnlp-main.240/.
- Fish, S., Gözl, P., Parkes, D. C., Procaccia, A. D., Rusak, G.,
Shapira, I., and Wüthrich, M. Generative social choice.
Journal of the ACM, 73(2):1–52, 2026.
- Fisher, J., Appel, R. E., Park, C. Y., Potter, Y., Jiang, L.,
Sorensen, T., Feng, S., Tsvetkov, Y., Roberts, M., Pan,
J., Song, D., and Choi, Y. Position: Political Neutrality
in AI Is Impossible — But Here Is How to Approximate
It. In *Forty-second International Conference on Machine*
Learning Position Paper Track, 2025. URL [https://](https://openreview.net/forum?id=H72JEXAPwo)
openreview.net/forum?id=H72JEXAPwo.
- Ge, L., Halpern, D., Micha, E., Procaccia, A. D., Shapira, I.,
Vorobeychik, Y., and Wu, J. Axioms for AI alignment
from human feedback. In Globerson, A., Mackey, L.,
Belgrave, D., Fan, A., Paquet, U., Tomczak, J., and
Zhang, C. (eds.), *Advances in Neural Information Pro-*
cessing Systems, volume 37, pp. 80439–80465. Curran
Associates, Inc., 2024. URL [https://proceedings.](https://proceedings.neurips.cc/paper_files/paper/2024/file/9328208f88ec69420031647e6ff97727-Paper-Conference.pdf)
[neurips.cc/paper_files/paper/2024/file/](https://proceedings.neurips.cc/paper_files/paper/2024/file/9328208f88ec69420031647e6ff97727-Paper-Conference.pdf)
9328208f88ec69420031647e6ff97727-Paper-Conference.
pdf.

- 495 Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian,
496 A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A.,
497 Vaughan, A., et al. The llama 3 herd of models. *arXiv*
498 *preprint arXiv:2407.21783*, 2024.
- 499
500 Hewitt, L., Ashokkumar, A., Ghezze, I., and Willer, R.
501 Predicting Results of Social Science Experiments using
502 Large Language Models. *Preprint*, 2024.
- 503
504 Hollender, A. and Kraiczy, S. Enforcing axioms for AI
505 alignment under loss-based rules. In *The Fourteenth*
506 *International Conference on Learning Representations*,
507 2026. URL <https://openreview.net/forum?id=MpYS0TK65s>.
- 508
509 Hsieh, C.-K., Yang, L., Cui, Y., Lin, T.-Y., Belongie, S., and
510 Estrin, D. Collaborative metric learning. In *Proceedings*
511 *of the 26th International Conference on World Wide Web*,
512 pp. 193–201, 2017. doi: 10.1145/3038912.3052639.
- 513
514 Hu, E. J., yelong shen, Wallis, P., Allen-Zhu, Z., Li,
515 Y., Wang, S., Wang, L., and Chen, W. LoRA: Low-
516 Rank Adaptation of Large Language Models. In *Inter-*
517 *national Conference on Learning Representations*,
518 2022. URL <https://openreview.net/forum?id=nZeVKeeFYf9>.
- 519
520 Hu, T., Baumann, J., Lupo, L., Collier, N., Hovy, D., and
521 Röttger, P. SimBench: Benchmarking the Ability of Large
522 Language Models to Simulate Human Behaviors. In *The*
523 *Fourteenth International Conference on Learning Rep-*
524 *resentations*, 2026. URL <https://openreview.net/forum?id=PL51SpN6ZJ>.
- 525
526
527 Jerrett, C. and Anshelevich, E. Low Cost, Fair, and Repre-
528 sentative Committees in a Metric Space. *arXiv preprint*
529 *arXiv:2510.21039*, 2025.
- 530
531 Kellerhals, L. and Peters, J. Proportional fairness
532 in clustering: A social choice perspective. In
533 Globerson, A., Mackey, L., Belgrave, D., Fan, A.,
534 Paquet, U., Tomczak, J., and Zhang, C. (eds.), *Ad-*
535 *vances in Neural Information Processing Systems*,
536 volume 37, pp. 111299–111317. Curran Asso-
537 ciates, Inc., 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/c981fd12b1d5703f19bd8289da9fc996-Paper-Conference.pdf.
- 538
539
540
541 Kolluri, A., Wu, S., Park, J. S., and Bernstein, M. S. Fine-
542 tuning LLMs for Human Behavior Prediction in So-
543 cial Science Experiments. In Christodoulopoulos, C.,
544 Chakraborty, T., Rose, C., and Peng, V. (eds.), *Pro-*
545 *ceedings of the 2025 Conference on Empirical Meth-*
546 *ods in Natural Language Processing*, pp. 30096–30111,
547 Suzhou, China, November 2025. Association for Com-
548 putational Linguistics. ISBN 979-8-89176-332-6. doi:
549 10.18653/v1/2025.emnlp-main.1530. URL <https://aclanthology.org/2025.emnlp-main.1530/>.
- Konya, A., Schirch, L., Irwin, C., and Ovadya, A. Demo-
cratic policy development using collective dialogues and
AI. *arXiv preprint arXiv:2311.02242*, 2023.
- Konya, A., Thorburn, L., Almasri, W., Leshem, O. A., Pro-
caccia, A., Schirch, L., and Bakker, M. Using collective
dialogues and AI to find common ground between Is-
raeli and Palestinian peacebuilders. In *Proceedings of the*
2025 ACM Conference on Fairness, Accountability, and
Transparency, FAccT ’25, pp. 312–333, New York, NY,
USA, 2025. Association for Computing Machinery. ISBN
9798400714825. doi: 10.1145/3715275.3732022. URL
<https://doi.org/10.1145/3715275.3732022>.
- Koren, Y., Bell, R., and Volinsky, C. Matrix factorization
techniques for recommender systems. *Computer*, 42(8):
30–37, 2009. doi: 10.1109/MC.2009.263.
- Kraiczy, S. and Elkind, E. Exact learning of preference
structure: Single-peaked preferences and beyond. In
Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu,
G., and Sabato, S. (eds.), *Proceedings of the 39th Interna-*
tional Conference on Machine Learning, volume 162 of
Proceedings of Machine Learning Research, pp. 11598–
11612. PMLR, 2022. URL <https://proceedings.mlr.press/v162/kraiczy22a.html>.
- Loshchilov, I. and Hutter, F. Decoupled Weight Decay
Regularization. In *International Conference on Learn-*
ing Representations, 2019. URL <https://openreview.net/forum?id=Bkg6RiCqY7>.
- Merrill, Samuel, I. and Grofman, B. *A Unified Theory*
of Voting: Directional and Proximity Spatial Models.
Cambridge University Press, Cambridge, 1999.
- Miyazaki, S. and Hall, A. B. Why Do AI Models Tell Left-
Wing Voters to Support the Communist Party? AI Voting
Advice in Japan’s 2026 General Election. 2026.
- Naous, T., Laban, P., Xu, W., and Neville, J. Flipping the Di-
alogue: Training and Evaluating User Language Models.
In *The Fourteenth International Conference on Learn-*
ing Representations, 2026. URL <https://openreview.net/forum?id=ykSmkVqzn4>.
- Peters, D. Recognising multidimensional euclidean prefer-
ences. In *Proceedings of the AAAI Conference on Artifi-*
cial Intelligence, volume 31, 2017.
- Poole, K. T. and Rosenthal, H. A spatial model for leg-
islative roll call analysis. *American Journal of Political*
Science, 29(2):357–384, 1985. doi: 10.2307/211172.

- 550 Robbins, H. A remark on Stirling’s formula. *The Amer-*
551 *ican Mathematical Monthly*, 62(1):26–29, 1955. doi:
552 10.2307/2308012. URL [https://doi.org/10.2307/](https://doi.org/10.2307/2308012)
553 [2308012](https://doi.org/10.2307/2308012).
- 554 Serfling, R. J. Probability inequalities for the sum in sam-
555 pling without replacement. *Annals of Statistics*, 2(1):
556 39–48, 1974. doi: 10.1214/aos/1176342611.
- 558 Sharma, M., McCain, M., Douglas, R., and Duvenaud, D.
559 Who’s in Charge? Disempowerment Patterns in Real-
560 World LLM Usage. *arXiv preprint arXiv:2601.19062*,
561 2026.
- 563 Sorensen, T., Moore, J., Fisher, J., Gordon, M., Miresghal-
564 lah, N., Rytting, C. M., Ye, A., Jiang, L., Lu, X., Dziri, N.,
565 Althoff, T., and Choi, Y. Position: A Roadmap to Plural-
566 istic Alignment. In *Proceedings of the 41st International*
567 *Conference on Machine Learning, ICML’24*. JMLR.org,
568 2024.
- 569 Wu, S., Choi, E., Khatua, A., Wang, Z., He-Yueya, J.,
570 Weerasooriya, T. C., Wei, W., Yang, D., Leskovec, J.,
571 and Zou, J. Humanlm: Simulating users with state
572 alignment beats response imitation. 2026. URL [https:](https://humanlm.stanford.edu/)
573 [//humanlm.stanford.edu/](https://humanlm.stanford.edu/).
- 575 Xu, A. and Davenport, M. Simultaneous preference and
576 metric learning from paired comparisons. In *Advances*
577 *in Neural Information Processing Systems*, volume 33,
578 2020.
- 580 Yao, S., Shinn, N., Razavi, P., and Narasimhan, K. R. τ -
581 bench: A benchmark for tool-agent-user interaction in
582 real-world domains. In *The Thirteenth International*
583 *Conference on Learning Representations*, 2025. URL
584 <https://openreview.net/forum?id=roNSXZpUDN>.
- 585 Zhang, Y., Li, M., Long, D., Zhang, X., Lin, H., Yang, B.,
586 Xie, P., Yang, A., Liu, D., Lin, J., Huang, F., and Zhou,
587 J. Qwen3 embedding: Advancing text embedding and
588 reranking through foundation models. *arXiv preprint*
589 *arXiv:2506.05176*, 2025.
- 591 Zhou, X., Sun, W., Ma, Q., Xie, Y., Liu, J., Du, W., Welleck,
592 S., Yang, Y., Neubig, G., Wu, S. T., and Sap, M. Mind
593 the Sim2Real Gap in User Simulation for Agentic Tasks,
594 2026. URL <https://arxiv.org/abs/2603.11245>.
- 595
596
597
598
599
600
601
602
603
604

A. Additional Details for Section 3

A.1. Viewpoint Embeddings and Related Work

The Viewpoint Embedding Space Hypothesis assumes the existence of a latent embedding space in which each agent is associated with an ideal point representing that agent’s viewpoint, and in which preferences are approximately spatial: each agent prefers viewpoints whose embeddings lie closer to their ideal point to those lying farther away. Crucially, distances in this space must reflect preferences—the closer a viewpoint is to an agent’s ideal point, the more that agent prefers it. Once relative distances encode comparative preference between viewpoints, one can reason geometrically about which groups of agents are represented to what degree by a slate of outputs. This, in turn, is what allows us to adapt tools from *proportional clustering* (Aziz et al., 2024; Chen et al., 2019) to our open-ended setting.

The viewpoint embedding hypothesis forms the basis of the spatial theory of voting (Bennett & Hays, 1960; Enelow & Hinich, 1984; Merrill & Grofman, 1999; Poole & Rosenthal, 1985). Social choice studies spatial restrictions such as single-peaked and Euclidean preferences, in which voters and alternatives are embedded on a latent axis or in a Euclidean space and preferences decrease with distance from each voter’s ideal point (Black, 1948; Bogomolnaia & Laslier, 2007; Elkind et al., 2022). This line of work is closely related to our viewpoint embedding hypothesis, but it typically assumes a finite set of alternatives and focuses on either recognizing whether an observed preference profile admits a spatial representation or recovering the underlying axis or preference structure (Peters, 2017; Kraiczy & Elkind, 2022). More recent work at the intersection of social choice and AI alignment adopts related shared-embedding assumptions, often modeling agents’ preferences with dot-product utilities (Ge et al., 2024; Hollender & Kraiczy, 2026)¹⁴.

The viewpoint embedding space assumption is closely related to embedding-based preference models, which learn low-dimensional representations of agents and alternatives in order to predict preferences. Latent-factor recommender systems typically score user–item pairs using interactions such as inner products (Koren et al., 2009). Inner-product latent factor models are closely related to ideal-point models: when user and item embeddings are constrained to the unit sphere, maximizing the user–item inner product is equivalent to minimizing Euclidean distance. Closest to our setting are metric-learning approaches, which give the embedding space an explicit geometric interpretation by learning representations in which smaller user–item distances correspond to stronger preferences (Hsieh et al., 2017; Xu & Davenport, 2020). Our viewpoint embedding hypothesis assumes an analogous distance-based structure, but over open-ended viewpoints rather than a fixed catalog of items.

Our approach does not require the hypothesis to hold exactly. What matters is not exact recovery of the true latent space, but that the learned embedding preserves the relevant preference geometry sufficiently well. This is important because, even if a viewpoint embedding space exists, it should not be expected to be directly observable. In Appendix F.3, we discuss how to learn a *viewpoint embedding function* ψ together with the agents’ viewpoint embeddings, and how this learned structure can be used to scale our method to inference-time.

A.2. Counterexample for k -means

While intuitive, neither PRF is not guaranteed by standard clustering techniques such as, for example, k -means or even balanced k -means, as the next example demonstrates.

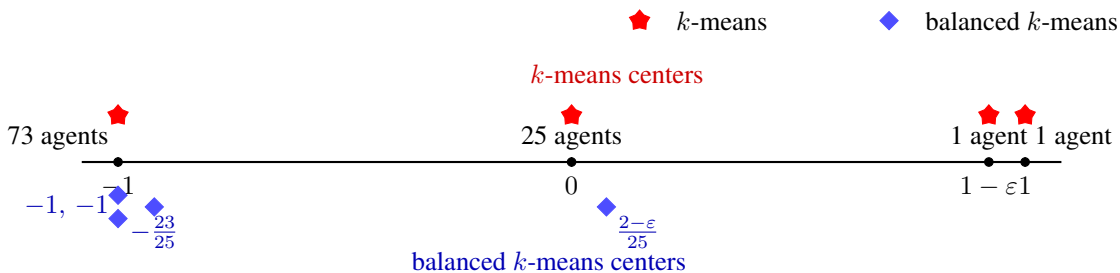


Figure A.1. The viewpoint embedding space from Example A.1, and outputs of (balanced) k -means for $k = 4$.

Example A.1. Suppose for simplicity that the viewpoint embedding space is the line segment $[-1, 1]$. There are $n = 100$

¹⁴These assumptions are closely connected to our distance-based formulation: on the unit sphere, maximizing dot-product utility is equivalent to minimizing Euclidean distance.

agents, consisting of four groups:

- Group 1: 73 agents at $x = -1$,
- Group 2: 25 agents at $x = 0$,
- Group 3: 1 agent at $x = 1 - \varepsilon$ for some small $\varepsilon > 0$, and
- Group 4: 1 agent at $x = 1$.

If we run k -means with $k = 4$ to cluster these agents, each group will be a cluster on its own. However, agents in Group 1 comprise more than half of the population, and therefore are entitled to $k/2 = 2$ of the responses. Since Group 1 only has one response in the output slate that is as close to them as their radius (in this case, $\Delta(S) = 0$), they are underrepresented, and the slate violates PRF (Definition 3.1) for $\ell = 2$.

On the other hand, if we use balanced k -means (so forcing each cluster to be of size $100/4 = 25$) we will end up with two clusters consisting entirely of agents in Group 1, one cluster with 23 agents from Group 1 and 2 agents from Group 2 (mean = $-\frac{23}{25}$), and one cluster with all remaining agents (mean = $\frac{2-\varepsilon}{25}$). However, agents in Group 2 make up one fourth of the population, and by PRF (for $\ell = 1$) they deserve influence over one out of the four responses. Since these agents would prefer a cluster consisting of just themselves (hence with mean $x = 0$, a response they all like the best) over any of the clusters output by balanced k -means, we conclude that Group 2 is indeed underrepresented.

A.3. Spatial Expanding Approvals Rule

In this section, we construct a faster implementation of SEAR. We first report the runtime shown in by Aziz et al. (Aziz et al., 2024).

Theorem A.2 ((Aziz et al., 2024)). *SEAR terminates in polynomial time $O(n^4 k^2)$.*

We first describe SEAR in pseudocode. Each agent $q \in \mathcal{N}$ has an initial budget $b_q = 1$. In each round, every point $p \in \mathcal{N}$ expands a ball around itself until the agents in the ball have total remaining budget at least n/k . The point whose ball reaches this threshold at the smallest radius is selected, and the agents in that ball pay a total of n/k .

The following theorem shows that Algorithm 1 gives a faster implementation of SEAR. We clearly remove the unnecessarily large dependency on n of n^4 by doing smart book-keeping, but additionally we are able to remove dependency on k entirely.

Theorem 3.2. *SEAR can be implemented with amortized running time $O(n^2 \log n)$.*

Proof. We describe an implementation of SEAR running in time $O(n^2 \log n)$.

First compute all pairwise distances and, for every $p \in \mathcal{N}$, sort the labels $\{1, \dots, n\}$ by increasing distance from p , breaking ties by label order. Let L_p denote this sorted list. This preprocessing takes $O(n^2 \log n)$ time.

The algorithm maintains the remaining budgets b_1, \dots, b_n . Initially $b_q = 1$ for all $q \in \mathcal{N}$. For each $p \in \mathcal{N}$, it also maintains an index $i(p)$ defined as the smallest index satisfying

$$\sum_{j=1}^{i(p)} b_{L_p[j]} \geq \frac{n}{k},$$

together with the partial sum

$$B(p) = \sum_{j=1}^{i(p)} b_{L_p[j]}.$$

Thus $\Delta(p, L_p[i(p)])$ is the current radius at which p 's expanding approval ball contains total remaining budget at least n/k .

In each round, the algorithm chooses

$$p^* \in \arg \min_{p \in \mathcal{N}} \Delta(p, L_p[i(p)])$$

715
716
717
718
719
720
721
722
723
724
725
726
727
728
729
730
731
732
733
734
735
736
737
738
739
740
741
742
743
744
745
746
747
748
749
750
751
752
753
754
755
756
757
758
759
760
761
762
763
764
765
766
767
768
769

Algorithm 1 Spatial Expanding Approvals Rule (SEAR)

Require: Agents $\mathcal{N} = (v_1, \dots, v_n)$, slate size k

- 1: $X \leftarrow ()$
- 2: $b_q \leftarrow 1$ for every $q \in \mathcal{N}$
- 3: **for** each $p \in \mathcal{N}$ **do**
- 4: Let $L_p = (L_p[1], \dots, L_p[n])$ be the agents ordered by increasing $\Delta(p, q)$, breaking ties by label order
- 5: **end for**
- 6: **for** $t = 1, \dots, k$ **do**
- 7: $A \leftarrow \{p \in \mathcal{N} : b_p > 0\}$
- 8: **for** each $p \in A$ **do**
- 9: Let $i(p)$ be the smallest index such that

$$\sum_{j=1}^{i(p)} b_{L_p[j]} \geq \frac{n}{k}.$$

- 10: $r_p \leftarrow \Delta(p, L_p[i(p)])$
 - 11: **end for**
 - 12: Choose $p^* \in \arg \min_{p \in \mathcal{A}} r_p$, breaking ties by label order
 - 13: Append v_{p^*} to X
 - 14: $\rho \leftarrow n/k$ {remaining amount to collect from the selected ball}
 - 15: **for** $j = 1, \dots, i(p^*)$ **do**
 - 16: $q \leftarrow L_{p^*}[j]$
 - 17: $x_q \leftarrow \min\{b_q, \rho\}$ {amount paid by agent q }
 - 18: $b_q \leftarrow b_q - x_q$ {update q 's remaining budget}
 - 19: $\rho \leftarrow \rho - x_q$ {update remaining joint payment}
 - 20: **if** $\rho = 0$ **then**
 - 21: **break**
 - 22: **end if**
 - 23: **end for**
 - 24: **end for**
 - 25: **return** W
-

and adds the point labelled p^* to W . It then charges the agents in

$$S = \{L_{p^*}[1], \dots, L_{p^*}[i(p^*)]\}$$

a total amount n/k : every agent in $S \setminus \{L_{p^*}[i(p^*)]\}$ spends its entire remaining budget, and the pivot agent $L_{p^*}[i(p^*)]$ pays just enough to bring the total payment to n/k . Since $i(p^*)$ is minimal, this is exactly the payment made by SEAR at the chosen radius.

We now bound the cost of maintaining the quantities $B(p)$ and $i(p)$. Suppose that an agent q pays amount $x_q > 0$. We subtract x_q from every partial sum $B(p')$ for which q appears among the first $i(p')$ entries of $L_{p'}$. This condition can be tested in $O(1)$ time using the distance matrix and the fixed tie-breaking rule: q appears at or before index $i(p')$ in $L_{p'}$ precisely when

$$(\Delta(p', q), q) \leq_{\text{lex}} (\Delta(p', L_{p'}[i(p')]), L_{p'}[i(p')]).$$

Hence processing one positive contribution costs $O(n)$, since we check all $p' \in \mathcal{N}$.

It remains to bound the total number of such positive contributions. Every non-pivot contributor spends its entire remaining budget, and therefore can be a non-pivot contributor at most once over the whole execution. Hence there are at most n positive non-pivot contributions in total. In addition, each selected point has at most one pivot contributor, so over the k rounds there are at most k pivot contributions. Therefore the total number of positive contributions is at most

$$n + k \leq 2n,$$

using $k \leq n$. Since each contribution costs $O(n)$ to propagate to all partial sums, all partial-sum updates take $O(n^2)$ time in total.

After these subtractions, some partial sums $B(p')$ may fall below n/k . For each $p' \in \mathcal{N}$, we restore the invariant by advancing $i(p')$ along $L_{p'}$ and adding the newly encountered budgets to $B(p')$ until

$$B(p') = \sum_{j=1}^{i(p')} b_{L_{p'}[j]} \geq \frac{n}{k}.$$

This takes $O(n)$ amortized time for each fixed p' , because $i(p')$ only increases and is bounded above by n . Thus the total time spent on threshold recomputation over all $p' \in \mathcal{N}$ and all rounds is $O(n^2)$.

Finally, in each round we find p^* by scanning all $p \in \mathcal{N}$, which costs $O(n)$ per round and hence $O(nk) \leq O(n^2)$ total. The preprocessing costs $O(n^2 \log n)$, while all subsequent scans, partial-sum updates, and threshold recomputations cost $O(n^2)$. Therefore the total running time is

$$O(n^2 \log n) + O(n^2) = O(n^2 \log n).$$

□

B. Computational Complexity of Verifying PRF

Given a possible outcome (either output by our algorithms or one of the approaches we are benchmarking against), we would like to check if the proportionality guarantees defined in the previous subsection are violated by this outcome. Without any restrictions (on the metric space), this problem is hard.

Theorem B.1. *Given a set of agents \mathcal{N} , a target output size k , and an outcome $X \subseteq \mathcal{M}$ with $|X| = k$, it is coNP-complete to check whether X satisfies PRF. This result holds even if the set of agents \mathcal{N} is in \mathbb{R}^d and the distance function is the Euclidean (ℓ_2) distance.*

Proof. Membership in coNP is straightforward: to show that a set X violates PRF, it suffices to guess a coalition of agents that witnesses this violation. To show coNP-hardness, we reduce from the classic CLIQUE problem. An instance of this problem is a graph $G = (V, E)$ together with a parameter t ; it is a yes-instance if G contains a clique of size t , and a no-instance otherwise. We will give two alternative proofs: a straightforward one using a general metric space, as well as a more involved proof using the Euclidean distance.

General metric space Given an instance $\langle G, t \rangle$ of CLIQUE, we construct an instance of unconstrained clustering as follows. We have two sets of agents:

- $\mathcal{N}_1 = V$. For every $a, b \in \mathcal{N}_1$, we set $\Delta(a, b) = 1$ if $(a, b) \in E$ and $\Delta(a, b) = 2$ otherwise;
- \mathcal{N}_2 is a set of size $|V|(t - 1)$. For every $a \in \mathcal{N}_2$ we set $\Delta(a, b) = 2$ for all $b \in \mathcal{N}_1 \cup \mathcal{N}_2$.

Set $k = |V|$, and let X be a subset of \mathcal{N}_2 of size k . Then:

1. d obeys triangle inequality since $\Delta(a, b) + \Delta(b, c) \geq 2 \geq \Delta(a, c)$ for any triple of agents $a, b, c \in \mathcal{N}$;
2. There can be no PRF violations for radius $y = 2$, since all agents are within distance 2 from each other;
3. There is a PRF violation for radius $y = 1$ if and only if there is a subset of \mathcal{N}_1 that is of size $n/k = t$ such that agents in that subset are within distance 1 from each other. This set corresponds to a clique in G .

Euclidean distance Consider an instance $\langle G, t \rangle$ of CLIQUE with $G = (V, E)$, where $|V| = n$. Label the vertices in V as $V = \{v_1, v_2, \dots, v_n\}$. Construct an instance of unconstrained clustering in \mathbb{R}^{n+1} as follows. Fix some $\varepsilon > 0$. We have two sets of agents:

- $\mathcal{N}_1 = \{c_i\}_{i \in [n]}$, i.e., we create one agent per vertex of the original graph. For each $i \in [n]$, we define the coordinates of $c_i \in \mathbb{R}^{n+1}$ as

$$(c_i)_j = \begin{cases} 1 & \text{if } j = i \\ \varepsilon & \text{if } j \leq n \text{ and } (v_i, v_j) \in E \\ 0 & \text{otherwise.} \end{cases}$$

- \mathcal{N}_2 is a set of size $n(t - 1)$. All $a \in \mathcal{N}_2$ have the same coordinates, defined as

$$(a)_j = \begin{cases} \sqrt{1 - 4\varepsilon + (n + 1)\varepsilon^2} & \text{if } j = n + 1 \\ 0 & \text{otherwise.} \end{cases}$$

Set $k = |V|$, and let X be a subset of \mathcal{N}_2 of size k .

For every $i \in [n]$ and $a \in \mathcal{N}_2$, we have

$$\Delta(c_i, a) = 1 - 4\varepsilon + (n + 1)\varepsilon^2 + 1 + \deg(v_i)\varepsilon^2 \in [2 - 4\varepsilon + (n + 1)\varepsilon^2, 2 - 4\varepsilon + 2n\varepsilon^2], \quad (1)$$

where $\Delta : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}_{\geq 0}$ is the squared Euclidean (ℓ_2) distance.

Given any $i, j \in [n]$ such that $(v_i, v_j) \in E$, we have

$$\Delta(c_i, c_j) \leq 2(1 - \varepsilon)^2 + (n - 2)\varepsilon^2 = 2 - 4\varepsilon + n\varepsilon^2. \quad (2)$$

Given any $i, j \in [n]$ such that $(v_i, v_j) \notin E$, we have

$$\Delta(c_i, c_j) \geq ((c_i)_i - (c_j)_i)^2 + ((c_i)_j - (c_j)_j)^2 = 2. \quad (3)$$

Let $\varepsilon = \frac{1}{n}$ and observe that

$$2 - 4\varepsilon + 2n\varepsilon^2 = 2 - \frac{4}{n} + \frac{2n}{n^2} = 2 - \frac{2}{n} < 2. \quad (4)$$

Observe that a subset of agents $S \subseteq \mathcal{N} = \mathcal{N}_1 \cup \mathcal{N}_2$ such that $S \cap \mathcal{N}_2 \neq \emptyset$ cannot witness a PRF violation: indeed, $\Delta(a, b) = 0$ for all $a \in \mathcal{N}_2$ and $b \in X$, so $|\{x \in X : \exists c \in S \text{ s.t. } \Delta(c, x) \leq y\}| = k$ for every $y \geq 0$.

We can therefore restrict our attention to coalitions $S \subseteq \mathcal{N}_1$. We claim that there exists such a coalition witnessing a PRF violation if and only if G has a clique of size t .

First consider the right to left direction. Given a clique $C \subseteq V$ of size $|C| = p$, let $S = \{c_i\}_{\{i \in [n]: v_i \in C\}}$ be the respective set of agents in \mathcal{N}_1 . We have $|S| = p = \frac{|\mathcal{N}_1|}{k}$. Then by (2), we have $\Delta(c_i, c_j) \leq 2 - 4\varepsilon + n\varepsilon^2$ for all $c_i, c_j \in S$. By (1), on the other hand, for every $c_i \in S$ and $x \in X$, it holds that $\Delta(c_i, x) \geq 2 - 4\varepsilon + (n+1)\varepsilon^2 > 2 - 4\varepsilon + n\varepsilon^2$. Thus, S witnesses a violation of PRF for $t = \sqrt{2 - 4\varepsilon + n\varepsilon^2}$.

Now consider the left to right direction. Let $S \subseteq \mathcal{N}$ be an ℓ -cohesive group that witnesses a violation of PRF for radius t . As shown above, we must have $S \subseteq \mathcal{N}_1$. Further, since S is ℓ -cohesive, we have $|S| \geq \frac{|\mathcal{N}_1|}{k} \geq \frac{|\mathcal{N}_1|}{k} = p$. By (1), we must have $t < 2 - 4\varepsilon + 2n\varepsilon^2$, since otherwise we would have had $|\{x \in X : \exists c \in S \text{ s.t. } \Delta(c, x) \leq t\}| = k$. Then for all $i, j \in [n]$ such that $c_i, c_j \in S$, we have $d^2(c_i, c_j) \leq t < 2 - 4\varepsilon + 2n\varepsilon^2 < 2$, where the last inequality follows from (4). By (3), this implies that $(v_i, v_j) \in E$. This proves that $\{v_i\}_{i \in [n]: c_i \in S}$ is a clique in G of size at least p . \square

C. Evaluation via Metric Extended Justified Representation+ (mEJR+)

Since it is hard to determine whether a given slate satisfies PRF, we now introduce another axiom, $mEJR+$, that (1) strengthens PRF and (2) is easy to verify. Further, as we will show, SEAR always outputs a slate that approximates $mEJR+$. Our goal is to use $mEJR+$ to evaluate the proportionality of the outcomes output by our method, as well as those output by methods we are comparing against.

Definition C.1 (Metric Extended Justified Representation (mEJR+)). An outcome $X \subseteq \mathcal{M}$ with $|X| = k$ satisfies Metric Extended Justified Representation+ (mEJR+) if there is no $c \in \mathcal{N}$, $t \geq 0$, and $S \subseteq \mathcal{N}$ such that S is ℓ -large and for all $i \in S$ we have

$$\Delta(i, c) \leq t \quad \text{and} \quad |\{c' \in X : \Delta(i, c') \leq t\}| < \ell.$$

To the best of our knowledge, $mEJR+$ has not been explicitly studied in the prior literature. However, it can be obtained by plugging the $EJR+$ axiom introduced by Brill & Peters (2023, Def. 10) for approval-based elections into the general framework introduced by Kellerhals & Peters (2024, Def. 4) for adapting approval-based proportionality axioms to the metric setting. Note that we explicitly restricted the candidate set to the set of agents: this is consistent with the PRF definition for unconstrained clustering as well as with how SEAR is defined.

First, we make a few observations about the $mEJR+$'s relationship to PRF (Definition 3.1) and SEAR (Appendix A.3). In what follows, we assume that $k < n$: this constraint is implicit in the definition of SEAR, as it picks the centers from the set of agents without replacement.

Proposition C.2. *Suppose that $k \leq n$. Then every slate that satisfies $mEJR+$ also satisfies PRF.*

Proof. Consider a slate $X \subseteq \mathcal{M}$ that violates PRF; we will argue that it violates $mEJR+$. Indeed, if X violates PRF, there exists an $\ell \in [k]$, $t > 0$, and a set of agents $S \subseteq \mathcal{N}$ of size at least $|S| \geq \ell n/k$ such that the maximum distance between pairs of agents in S is t , but

$$|\{c \in X : \exists i \in S \text{ s.t. } \Delta(i, c) \leq t\}| < \ell.$$

Then the set $S \setminus X$ is non-empty: otherwise, each agent $i \in S$ would have at least $\ell n/k \geq \ell$ outcomes in X at distance at most t from it, a contradiction. Let c be some agent in $S \setminus X$. Then, for each $i \in S$ we have $\Delta(i, c) \leq t$ by assumption, and also

$$|\{c' \in X : \Delta(i, c') \leq t\}| \leq |\{c' \in X : \exists i' \in S \text{ s.t. } \Delta(i', c') \leq t\}| < \ell.$$

Thus, S is a witness that X violates $mEJR+$. \square

Proposition C.3. *Let X be a slate output by SEAR (Appendix A.3). Then X is a 3-approximation for $mEJR+$ in the following sense: there is no agent $c \in \mathcal{N}$, group of agents $S \subseteq \mathcal{N}$, $t > 0$, and ℓ with $|S| \geq \ell n/k$ such that for all $i \in S$ it holds that*

$$\Delta(i, c) \leq y \quad \text{and} \quad |\{c' \in X : \Delta(i, c') \leq 3t\}| < \ell.$$

Proof. Suppose for contradiction that there is such a $c \in \mathcal{N}$ and a group of agents $S \subseteq \mathcal{N}$ as described in the proposition statement. Fix any $i \in S$. By the triangle inequality, for any $c' \in X$ such that $\Delta(i', c') \leq t$ for some $i' \in S$, we have $\Delta(i, c') \leq \Delta(i, c) + \Delta(c, i') + \Delta(i', c') \leq 3t$. Therefore we must have

$$|\{c' \in X : \exists i' \in S \text{ s.t. } \Delta(i', c') \leq t\}| \leq |\{c' \in X : \Delta(i, c') \leq 3t\}| < \ell.$$

Consider the snapshot of the SEAR algorithm when neighborhood distance t (or the first valid distance that is at least t) is considered. At this point, the agents in S have only used their budget/weight towards selecting outcomes less than t distance from them, so their total remaining budget is at least $\ell \frac{n}{k} - (\ell - 1) \frac{n}{k} = \frac{n}{k}$. So agents in S still have sufficient total weight/budget to select one more location that is a distance at most t from some agent in S , and in particular $c \in \mathcal{N}$ is at most t from each $i \in S$, so the algorithm would have selected at least one more center within distance t to some $i \in S$. \square

C.1. Efficiently Computing mEJR+ Approximation

Unlike PRF, which is hard to verify (Theorem B.1), mEJR+ can be verified efficiently. Not only can we improve substantially over the naive polynomial time verification with runtime $O(n^4)$ ¹⁵, but we show that further we can compute the *mEJR+ approximation*, defined in this section, just as efficiently in time $O(n^2(\log n + k))$ Algorithm 2. Such an approximation is much more useful than a binary satisfaction criterion, as it can be used as a fine-grained measure to compare slates produced by other algorithms, as we do in our experiments.

Let $n = |\mathcal{N}|$ and $[k] = \{1, \dots, k\}$ if k is a positive integer. As before, let (\mathcal{M}, Δ) be a metric space.

Definition C.4 (distance-approximate mEJR). Let $X \subseteq \mathcal{M}$ be a finite set of size k , where k is positive and $k \leq n$. Then X is a β -approximation for mEJR+ in the following sense: there is no agent $c \in \mathcal{N}$, group of agents $S \subseteq \mathcal{N}$, that is t -cohesive for $t \geq 0$, and $\ell \in [k]$ with $|S| \geq \ell n/k$ such that

$$|\{c' \in X : \Delta(i, c') \leq \beta t\}| < \ell$$

Thus if X satisfies mEJR+ if and only if it is a β -approximation of mEJR+ for $\beta \leq 1$. We remark that the second condition in Definition C.4 can be equivalently written as

$$\Delta_\ell(i, X) > \beta t,$$

where $\Delta_\ell(i, X)$ denotes the distance from i to its ℓ th closest element in X (with ties broken arbitrarily).

Let β_{opt} be the approximation ratio of slate X on the instance \mathcal{N} , i.e. the smallest value of β such that X is a β -distance-approximation of mEJR+ in the sense of Definition C.4. Let $X \subseteq \mathcal{M}$ with $|X| = k$. Let $\Delta_\ell(i, X)$ be the distance from i to the ℓ th closest element in X . Let

$$\beta_{opt}^\ell(S) = \frac{\min_{i \in S} \Delta_\ell(i, X)}{\min_{c \in \mathcal{N}} \max_{i \in S} \Delta(i, c)}.$$

and

$$\beta_{opt}^\ell = \max_{S \subseteq \mathcal{N}, |S| \geq \ell n/k} \beta_{opt}^\ell(S).$$

So $\beta_{opt} = \max_\ell \beta_{opt}^\ell$. Let β be the approximation ratio output by Algorithm 2 and let β^ℓ be the approximation ratio computed at the end of the ℓ th outer loop of Algorithm 2. We will show that $\beta^\ell = \beta_{opt}^\ell$ from which it follows that $\beta = \beta_{opt}$.

Remark C.5 (Invariant for i_c). Throughout the inner loop (for fixed ℓ), for each $c \in \mathcal{N}$ let $q := \lceil \ell n/k \rceil$. The pointer i_c is maintained so that $L_c[i_c]$ is the q th active agent in the list L_c (and $i_c = \text{None}$ if fewer than q active agents remain).

Lemma C.6 (Runtime). *The runtime is of Algorithm 2 $O(n^2(\log n + k))$.*

Proof. We first analyze the preprocessing. For each $c \in \mathcal{N}$, we construct the list L_c of agents in \mathcal{N} sorted by increasing distance from c . This takes $O(n^2 \log n)$ time over all c . Similarly, for each $u \in \mathcal{N}$, we construct the list X_u of elements of X sorted by increasing distance from u , which takes $O(nk \log k)$ time. Since $k \leq n$, the total preprocessing time is

$$O(n^2 \log n + nk \log k) = O(n^2 \log n).$$

¹⁵Tractability by itself is not surprising: since EJR+ can be checked efficiently in $O(n^2)$ time in the approval setting (when the candidates are the set of agents), we could naively check for EJR+ violations for the approval instances induced by every possible value of y , of which there are $O(n^2)$ many (the distance for each possible pair of agents). However, this gives an overall runtime of $O(n^4)$ which can still be prohibitive in settings with large n , such as ours.

Algorithm 2 mEJR+ Approximation

Require: agents \mathcal{N} in metric space (\mathcal{M}, Δ) , slate X of size k
 1: For each $c \in \mathcal{N}$, create list L_c of agents in \mathcal{N} sorted by dist from c *{Preprocessing}*
 2: For each $u \in \mathcal{N}$, create list X_u of agents in X sorted by dist from u
 3: $\beta \leftarrow 0$
 4: **for** $\ell = 1, \dots, k$ **do** *{Approximation Ratio}*
 5: $T \leftarrow$ all agents $u \in \mathcal{N}$ sorted by $\Delta_\ell(u, X) := \Delta(u, X_u[\ell])$
 6: Mark all agents in T as *active*
 7: $i_T \leftarrow 1$
 8: $\beta_\ell \leftarrow 0$
 9: $i_c \leftarrow \lceil \frac{\ell n}{k} \rceil$ for all c
 10: **while** $i_T \leq |\mathcal{N}|$ **do**
 11: $C_{\text{valid}} \leftarrow \{c \in \mathcal{N} \mid i_c \neq \text{None}\}$
 12: **if** $C_{\text{valid}} = \emptyset$ **then break**
 13: **end if**
 14: $r^* \leftarrow \min\{\Delta(c, L_c[i_c]) \mid c \in C_{\text{valid}}\}$
 15: $\beta_\ell \leftarrow \max(\beta_\ell, \frac{\Delta_\ell(T[i_T], X)}{r^*})$ **if** $r^* > 0$ **else return** ∞
 16: **for each** $c \in C_{\text{valid}}$ **do**
 17: **if** agent $T[i_T]$ has index j with $j \leq i_c$ in L_c **then**
 18: $i_c \leftarrow$ index ($> i_c$) of next *active* agent in L_c **else None**
 19: **end if**
 20: **end for**
 21: Mark agent $T[i_T]$ as *inactive*
 22: $i_T += 1$
 23: **end while**
 24: $\beta \leftarrow \max(\beta_\ell, \beta)$
 25: **end for**
 26: **return** β

Now fix an iteration of the outer loop, corresponding to some $\ell \in \{1, \dots, k\}$, and let $q = \lceil \frac{\ell n}{k} \rceil$. The algorithm first sorts the agents $u \in \mathcal{N}$ by

$$\Delta_\ell(u, X) = \Delta(u, X_u[\ell]).$$

Since the lists X_u have already been sorted, each value $\Delta_\ell(u, X)$ can be read in $O(1)$ time. Thus constructing and sorting T takes $O(n \log n)$ time. Marking all agents as active, initializing i_T, β_ℓ , and setting $i_c \leftarrow q$ for every $c \in \mathcal{N}$ takes $O(n)$ additional time.

We next bound the total time spent in the while loop for this fixed ℓ . The pointer i_T increases by one in every non-terminal iteration, and therefore the while loop has at most n non-terminal iterations.

In each iteration, the set

$$C_{\text{valid}} = \{c \in \mathcal{N} : i_c \neq \text{None}\}$$

can be found by scanning all $c \in \mathcal{N}$, taking $O(n)$ time. The value

$$r^* = \min\{\Delta(c, L_c[i_c]) : c \in C_{\text{valid}}\}$$

can also be computed by scanning all valid centers, again taking $O(n)$ time. Therefore these operations contribute $O(n^2)$ total time over the while loop.

It remains to bound the cost of updating the pointers i_c . For each $c \in C_{\text{valid}}$, we need to test whether the newly deactivated agent $T[i_T]$ appears at or before position i_c in L_c . This can be tested in $O(1)$ time, either by using precomputed ranks in the lists L_c , or equivalently by comparing the lexicographic distance-order keys. Hence these membership tests cost $O(n)$ per while-loop iteration and $O(n^2)$ total for this fixed ℓ .

Whenever the test succeeds for some c , the pointer i_c is advanced to the next active agent after its current position in L_c , or set to None if no such agent exists. For each fixed c , the pointer i_c only moves forward through the list L_c . Hence, over all

1045 iterations of the while loop, i_c advances at most n times. Summing over all $c \in \mathcal{N}$, the total time spent advancing pointers is
 1046 $O(n^2)$ for this fixed ℓ .

1047 Thus, for a fixed ℓ , the while loop costs $O(n^2)$ time, and the sorting and initialization before the while loop cost $O(n \log n)$.
 1048 Therefore one outer-loop iteration costs

$$1049 \quad O(n^2 + n \log n) = O(n^2).$$

1050 There are k iterations of the outer loop, so the total time after preprocessing is $O(kn^2)$.

1051 Combining this with the preprocessing cost, the total running time is

$$1052 \quad O(n^2 \log n) + O(kn^2) = O(n^2(\log n + k)).$$

1055 \square

1056 **Lemma C.7.** *If β^ℓ gets updated to β_j^ℓ during the j th round of the inner loop of Algorithm 2, then $\beta_{opt}^\ell \geq \beta_j^\ell$.*

1057 *Proof.* Consider the j th execution of the inner loop of Algorithm 2 and suppose in that round $c_j \in C_{valid}$ minimizes
 1058 $\Delta(c, L_c[i_c])$ over all $c \in C_{valid}$. As in Algorithm 2, let $r^* = \Delta(c_j, L_{c_j}[i_{c_j}])$. Let $q := \lceil \ell n/k \rceil$ and let S be the set of the q
 1059 active agents among the first i_{c_j} positions in L_{c_j} . Define $j' \geq j$ to be the first round (counting from round j onward) in
 1060 which some agent from S becomes inactive, and let u be that agent. Then at the beginning of round j' all agents in S are
 1061 active, and since agents become inactive in nondecreasing order of $\Delta_\ell(\cdot, X)$ (according to T), we have

$$1062 \quad \Delta_\ell(u, X) = \min_{i \in S} \Delta_\ell(i, X).$$

1063 Moreover, in round j we have $\Delta(i, c_j) \leq r^*$ for all $i \in S$ (by definition of r^* and S), hence $\max_{i \in S} \Delta(i, c_j) \leq r^*$ and
 1064 $\min_{i \in S} \Delta_\ell(i, X) = \Delta_\ell(u, X)$. Therefore

$$1065 \quad \beta_{opt}^\ell \geq \frac{\min_{i \in S} \Delta_\ell(i, X)}{\max_{i \in S} \Delta(i, c_j)} \geq \frac{\Delta_\ell(u, X)}{r^*},$$

1066 and since $T[i_T]$ is the first active agent in T in round j , $\Delta_\ell(u, X) \geq \Delta_\ell(T[i_T], X)$, yielding

$$1067 \quad \beta_{opt}^\ell \geq \frac{\Delta_\ell(u, X)}{r^*} \geq \frac{\Delta_\ell(T[i_T], X)}{r^*} = \beta_j^\ell.$$

1074 \square

1075 **Lemma C.8 (No-skipping).** *For any group S with $|S| \geq \ell n/k$, $\beta_{opt}^\ell(S) \leq \beta^\ell$.*

1076 *Proof.* Fix S with $|S| \geq \ell n/k$ and consider the ℓ th outer iteration of Algorithm 2. Let u be the first agent in S that becomes
 1077 inactive; then $u = T[i]$ for the minimum i with $T[i] \in S$, and hence

$$1078 \quad \Delta_\ell(u, X) = \min_{i \in S} \Delta_\ell(i, X).$$

1079 In the round when u becomes inactive, β^ℓ is updated to at least $\Delta_\ell(u, X)/r^*$, where $r^* = \min\{\Delta(c, L_c[i_c]) : c \in C_{valid}\}$.
 1080 Let c^* minimize $\max_{i \in S} \Delta(i, c)$ over $c \in \mathcal{N}$. Before removing u , all agents in S are active, so $c^* \in C_{valid}$ and (by the
 1081 invariant for i_{c^*}) $\Delta(c^*, L_{c^*}[i_{c^*}]) \leq \max_{i \in S} \Delta(i, c^*)$ in that round. Therefore

$$1082 \quad r^* \leq \Delta(c^*, L_{c^*}[i_{c^*}]) \leq \max_{i \in S} \Delta(i, c^*) = \min_{c \in \mathcal{N}} \max_{i \in S} \Delta(i, c).$$

1083 Plugging into the update yields

$$1084 \quad \beta^\ell \geq \frac{\Delta_\ell(u, X)}{r^*} \geq \frac{\min_{i \in S} \Delta_\ell(i, X)}{\min_{c \in \mathcal{N}} \max_{i \in S} \Delta(i, c)} = \beta_{opt}^\ell(S),$$

1085 as claimed. \square

1086 **Theorem C.9.** *Algorithm 2 computes the optimal approximation ratio of a slate X for instance \mathcal{N} , i.e.*

$$1087 \quad \beta = \beta_{opt}.$$

1088 *Proof.* By Lemma C.7, it holds that $\beta_{opt}^\ell \geq \beta_j^\ell$ whenever β^ℓ gets updated to β_j^ℓ in the j th round of the ℓ th outer loop.
 1089 Thus, $\beta^\ell = \max_j \beta_j^\ell \leq \beta_{opt}^\ell$. It follows that $\beta_{opt} := \max_\ell \beta_{opt}^\ell \geq \max_\ell \beta^\ell =: \beta$. For the reverse inequality, note that
 1090 $\beta_{opt}^\ell = \max_{S \subseteq \mathcal{N}, |S| \geq \ell n/k} \beta_{opt}^\ell(S)$, by Lemma C.8, $\beta_{opt}^\ell \leq \beta^\ell$. It follows that $\beta_{opt} \leq \beta$. \square

D. Additional Details for Section 4

Theorem 4.2 (Amortized Selection–Generation Commutation). *Let $\{\phi_i(p)\}_{i \in \mathcal{N}}$ be γ -proxy viewpoint embeddings. For every ℓ -large and t -cohesive group of agents $S \subseteq \mathcal{N}$ there exist at least ℓ outputs $o \in Y$ such that for each such o , there exists some agent $i \in S$ satisfying*

$$\Delta(\psi(o), \psi(o_i^*)) \leq t + 2\gamma.$$

Proof. Select-then-Generate returns a slate of responses $Y = \{A_i(p) \mid i \in R\}$. Each $o \in Y$ is generated by using a simulation query from agent i 's simulation on prompt p and this $i \in R$ corresponds to a viewpoint in X selected by SEAR run when on the proxy embeddings $\{\phi_i(p)\}$. We will show that PRF over the approximate viewpoint embeddings, the condition implied by SEAR, implies approximate PRF with respect to the true viewpoint embeddings $\{\psi(A_i(p))\}_{i \in \mathcal{N}}$. Let S be an ℓ -large and t -cohesive group. By definition of γ -proxy embeddings, we have that for any $x \in X$

$$\Delta(\phi_x(p), \psi(A_x(p))) \leq \gamma.$$

If in addition for some $i \in S$ we have

$$\Delta(\phi_i(p), \phi_x(p)) \leq t,$$

then by the triangle inequality it follows that $\Delta(\phi_i(p), \psi(A_x(p))) \leq \gamma + t$. Therefore, the PRF representation guarantee that $|\{x \in X : \exists i \in S \text{ s.t. } \Delta(\phi_i(p), \phi_x(p)) \leq t\}| \geq \ell$ translates to

$$\begin{aligned} \ell &\leq |\{x \in R : \exists i \in S \text{ s.t. } \Delta(\phi_i(p), \phi_x(p)) \leq t\}| \\ &\leq |\{x \in R : \exists i \in S \text{ s.t. } \Delta(\phi_i(p), \psi(A_x(p))) \leq t + \gamma\}| \\ &\leq |\{x \in R : \exists i \in S \text{ s.t. } \Delta(\psi(A_i(p)), \psi(A_x(p))) \leq t + 2\gamma\}| \\ &= |\{o \in Y : \exists i \in S \text{ s.t. } \Delta(\psi(o_i^*), \psi(o)) \leq t + 2\gamma\}|. \end{aligned}$$

as claimed. □

E. Sampling Approximate PRF slates

E.1. Approximate Proportionality Guarantees for uniform sampling without replacement

Theorem 4.3 (Population Guarantees from Sampled Agents). *Let \mathcal{M} denote the underlying population, with $|\mathcal{M}| = m$. Let $\mathcal{N} \subseteq \mathcal{M}$ be a sample of n agents drawn uniformly at random without replacement. If $n = \Omega\left(\frac{1}{\varepsilon^2} \log \frac{km}{\delta}\right)$ and Y^* is a PRF slate with respect to the sampled agents \mathcal{N} , then with probability at least $1 - \delta$, for every $\ell \in [k]$ and every t -cohesive group $S \subseteq \mathcal{M}$ with $|S| \geq (\ell/k + \varepsilon)m$,*

$$|\{o \in Y^* \mid \exists i \in S \text{ s.t. } \Delta(\psi(o_i^*), \psi(o)) \leq 2t\}| \geq \ell.$$

Proof. Consider kn sets defined as follows:

For each $v \in \mathcal{N}$ and $\ell \in [k]$, let S_v^ℓ be a subset of \mathcal{N} of size $\lceil (\ell/k + \varepsilon)n \rceil$ that contains v and minimizes diameter among all subsets of \mathcal{N} of that size containing v .

Let \mathcal{N}' be a sample of \mathcal{N} with $|\mathcal{N}'| = d$. For any $\ell \in [k]$, define $S'^\ell = \mathcal{N}' \cap S_v^\ell$ and let $\hat{p}_v^\ell = \frac{|S'^\ell|}{d}$.

Theorem E.1 (Serfling–Hoeffding, (Serfling, 1974)). *Let $a_1, \dots, a_N \in [A, B]$ with mean $\mu = \frac{1}{N} \sum_{i=1}^N a_i$. Sample without replacement Y_1, \dots, Y_n and set $\bar{Y}_n = \frac{1}{n} \sum_{j=1}^n Y_j$. Then for any $t > 0$,*

$$\Pr(|\bar{Y}_n - \mu| \geq t) \leq 2 \exp\left(-\frac{2nt^2}{\left(1 - \frac{n-1}{N}\right)(B-A)^2}\right).$$

Using a Serfling-Hoeffding bound,

$$\begin{aligned} \Pr\left(\frac{\ell}{k} > \hat{p}_v^\ell\right) &= \Pr\left(\frac{\ell}{k} + \varepsilon - \hat{p}_v^\ell > \varepsilon\right) \leq \Pr(|\hat{p}_v^\ell - \mathbb{E}[\hat{p}_v^\ell]| > \varepsilon) \leq 2 \exp\left(-\frac{2d\varepsilon^2}{(1 - \frac{d-1}{n})}\right) \\ &\leq 2 \exp(-2d\varepsilon^2). \end{aligned}$$

Union bounding over all kn sets we get that

$$\Pr\left(\max_{i \in \mathcal{N}, \ell \in [k]} \frac{\ell}{k} - \hat{p}_i^\ell > 0\right) \leq 2kn \exp(-2d\varepsilon^2),$$

thus taking

$$d \geq \frac{1}{2\varepsilon^2} \log\left(\frac{2kn}{\delta}\right)$$

suffices so that w.p. at least $1 - \delta$, for all $i \in \mathcal{N}$ and $\ell \in [k]$, $\hat{p}_i^\ell \geq \frac{\ell}{k}$.

Consider any set S of agents of size at least $(\frac{\ell}{k} + \varepsilon)n$ where $\ell \in [k]$ who all have pairwise distance at most t . Pick an arbitrary agent $v \in S$. Since S contains at least $\lceil (\ell/k + \varepsilon)n \rceil$ agents, includes v , and has diameter at most t , it contains a feasible set for the definition of S_v^ℓ with diameter at most t . By minimality, $\text{diam}(S_v^\ell) \leq t$, and hence $\text{diam}(S'^\ell_v) \leq t$.

Next, observe that if we run SEAR on \mathcal{N}' , then by Theorem F.2 there are at least ℓ candidates in W^* that are at distance t to some member of S'^ℓ_v . Since every $i \in S'^\ell_v \subseteq S_v^\ell$ has distance at most t to $v \in S$, we conclude that there exist ℓ points in W^* at distance at most $2t$ to v . This concludes the proof. \square

E.2. Stratified Sampling without Replacement

Suppose for $t \geq 1$, a constant, we have a partition (P_1, \dots, P_t) of \mathcal{M} such that $|P_i| = r_i |\mathcal{N}|$ where $\sum_i r_i = 1$ with $\min_i r_i > 0$. Let $m := |\mathcal{M}|$.

Definition E.2. We say that a subset \mathcal{N} of \mathcal{M} is stratified if

$$|\mathcal{N} \cap P_i| \in \{\lceil r_i |\mathcal{N}| \rceil, \lfloor r_i |\mathcal{N}| \rfloor\} \quad \text{for all } 1 \leq i \leq t.$$

Theorem E.3. Let $\varepsilon, \delta > 0$. Let \mathcal{N} be sampled uniformly at random without replacement among stratified samples from \mathcal{M} of size n , with

$$n = \Omega\left(\frac{1}{\varepsilon^2} \left[\log \frac{n}{\delta} + t \log\left(\frac{t}{\varepsilon^2} \log \frac{n}{\delta}\right) \right]\right).$$

If W^* is a PRF slate with respect to S , then W^* is an up to one 2-approximate PRF slate with respect to \mathcal{N} with probability at least $1 - \delta$.

Proof. We will prove the result by using Theorem 4.3 in a black-box fashion.

We say that $S \subseteq \mathcal{N}$ is *bad* if a PRF slate for S is not up to one 2-approximate PRF for \mathcal{N} . Let B be the event that S is bad. Let A be the event that a set S is stratified with respect to \mathcal{N} (and the partition (P_1, \dots, P_t)). Set $n_i := |P_i|$ so that $\sum_i n_i = n$ and $r_i = n_i/n$. For a draw of size s , write feasible count vectors as

$$\mathcal{S} = \left\{ \mathbf{c} = (c_1, \dots, c_t) \in \mathbb{Z}_{\geq 0}^t : \sum_i c_i = s, 0 \leq c_i \leq n_i \right\}.$$

Let $\mathcal{C} \subseteq \mathcal{S}$ be the set of *stratified* vectors: $c_i \in \{\lfloor sr_i \rfloor, \lceil sr_i \rceil\}$ for all i .

Define the *multivariate hypergeometric* ($H(\mathbf{c})$, sampling without replacement) and *multinomial* ($B(\mathbf{c})$, sampling with replacement) pmfs on \mathcal{S} :

$$H(\mathbf{c}) = \frac{\prod_{i=1}^t \binom{n_i}{c_i}}{\binom{n}{s}}, \quad B(\mathbf{c}) = \frac{s!}{\prod_{i=1}^t c_i!} \prod_{i=1}^t r_i^{c_i}.$$

We claim that

$$\Pr(A) \geq H(\mathbf{c}^*) \geq B(\mathbf{c}^*) \geq \Omega\left(\frac{1}{s^t}\right),$$

for some stratified vector \mathbf{c}^* , where $s = |S|$ is the sample size under consideration and is large enough.

Claim E.4. Fix $t \geq 2$ and $\rho \in (0, 1)$ and assume $r_i := n_i/n \geq \rho$ for all i . Let $\mathbf{c} = (c_1, \dots, c_t)$ satisfy

$$c_i \in \{\lfloor sr_i \rfloor, \lceil sr_i \rceil\} \quad \text{for all } i,$$

where $s = \sum_i c_i \in \{1, \dots, n-1\}$. Then, for all s with $\min\{s, n-s\} \geq 2/\rho$,

$$H(\mathbf{c}) \geq Z'_t s^{-\frac{t-1}{2}}, \quad Z'_t := (2\pi)^{\frac{1-t}{2}} \rho^{t/2} \exp\left(-\frac{t}{6} - \frac{1}{12}\right).$$

For $\min\{s, n-s\} < 2/\rho$ one has the trivial bound $H(\mathbf{c}) \geq (s+1)^{1-t}$.

Proof. Recall

$$H(\mathbf{c}) = \frac{\prod_{i=1}^t \binom{n_i}{c_i}}{\binom{n}{s}} = \frac{\prod_{i=1}^t \frac{n_i!}{c_i!(n_i-c_i)!}}{\frac{n!}{s!(n-s)!}} = \frac{\left(\prod_i n_i!\right) s!(n-s)!}{n! \left(\prod_i c_i!\right) \left(\prod_i (n_i-c_i)!\right)}. \quad (\dagger)$$

We use Robbins' version of Stirling's bounds (Robbins, 1955): for every integer $m \geq 1$,

$$\sqrt{2\pi m} \left(\frac{m}{e}\right)^m \leq m! \leq \sqrt{2\pi m} \left(\frac{m}{e}\right)^m e^{1/(12m)}.$$

Apply the *lower* bound to $n_i!$, $s!$, $(n-s)!$ and the *upper* bound to $n!$, $c_i!$, $(n_i-c_i)!$ in (\dagger) . Plugging in and cancelling the common $(m/e)^m$ -terms, we obtain

$$H(\mathbf{c}) \geq (2\pi)^{\frac{1-t}{2}} \frac{\sqrt{s} \sqrt{n-s} \prod_i \sqrt{n_i}}{\sqrt{n} \prod_i \sqrt{c_i} \prod_i \sqrt{n_i-c_i}} \exp(-E_t), \quad (*)$$

where the error exponent is

$$E_t := \frac{1}{12n} + \sum_{i=1}^t \left(\frac{1}{12c_i} + \frac{1}{12(n_i-c_i)} \right).$$

We first lower-bound

$$R := \frac{\sqrt{s} \sqrt{n-s} \prod_i \sqrt{n_i}}{\sqrt{n} \prod_i \sqrt{c_i} \prod_i \sqrt{n_i-c_i}}.$$

Use the assumptions:

$$r_i = \frac{n_i}{n} \geq \rho, \quad \sum_i n_i = n, \quad \sum_i c_i = s.$$

For the numerator,

$$\prod_i \sqrt{n_i} = \sqrt{\prod_i n_i} \geq \sqrt{(n\rho)^t} = n^{t/2} \rho^{t/2}.$$

For the denominators, we use only crude upper bounds; Because $c_i \leq s$ and $\sum_i c_i = s$, we have $\prod_i \sqrt{c_i} \leq s^{t/2}$. Because $0 \leq n_i - c_i$ and $\sum_i (n_i - c_i) = n - s$, each $n_i - c_i \leq n - s$, so $\prod_i \sqrt{n_i - c_i} \leq (n-s)^{t/2}$.

Thus

$$R \geq \frac{\sqrt{s} \sqrt{n-s} n^{t/2} \rho^{t/2}}{\sqrt{n} s^{t/2} (n-s)^{t/2}} = \rho^{t/2} \frac{n^{\frac{t-1}{2}}}{s^{\frac{t-1}{2}} (n-s)^{\frac{t-1}{2}}}.$$

Since $n-s \leq n$, we have $n^{(t-1)/2} / (n-s)^{(t-1)/2} \geq 1$, hence $R \geq \rho^{t/2} s^{-\frac{t-1}{2}}$.

We now impose $\min\{s, n-s\} \geq 2/\rho$ as in the statement. For a stratified vector $c_i \in \{\lfloor sr_i \rfloor, \lceil sr_i \rceil\}$ with $r_i \geq \rho$, we have

$$c_i \geq sr_i - 1 \geq s\rho - 1 \geq \frac{1}{2}s\rho,$$

because $s\rho \geq 2$. Similarly,

$$n_i - c_i = nr_i - c_i \geq nr_i - (sr_i + 1) = (n-s)r_i - 1 \geq (n-s)\rho - 1 \geq \frac{1}{2}(n-s)\rho,$$

since $(n-s)\rho \geq 2$ under the same assumption.

Therefore,

$$\frac{1}{12c_i} \leq \frac{1}{6s\rho}, \quad \frac{1}{12(n_i - c_i)} \leq \frac{1}{6(n-s)\rho},$$

and hence

$$\sum_{i=1}^t \left(\frac{1}{12c_i} + \frac{1}{12(n_i - c_i)} \right) \leq \frac{t}{6s\rho} + \frac{t}{6(n-s)\rho} \leq \frac{t}{3 \min\{s, n-s\} \rho}.$$

Using $\min\{s, n-s\} \geq 2/\rho$ gives

$$\frac{t}{3 \min\{s, n-s\} \rho} \leq \frac{t}{3 \cdot (2/\rho) \cdot \rho} = \frac{t}{6}.$$

Finally, since $n \geq s \geq 2/\rho$, we also have $1/(12n) \leq 1/12$. Altogether,

$$E_t = \frac{1}{12n} + \sum_{i=1}^t \left(\frac{1}{12c_i} + \frac{1}{12(n_i - c_i)} \right) \leq \frac{1}{12} + \frac{t}{6}. \quad (***)$$

Thus

$$\exp(-E_t) \geq \exp\left(-\frac{t}{6} - \frac{1}{12}\right).$$

Substituting (**) and (***) into (*) yields

$$H(\mathbf{c}) \geq (2\pi)^{\frac{1-t}{2}} R \exp(-E_t) \geq (2\pi)^{\frac{1-t}{2}} \rho^{t/2} \exp\left(-\frac{t}{6} - \frac{1}{12}\right) s^{-\frac{t-1}{2}}.$$

This is precisely the claimed bound with

$$Z'_t := (2\pi)^{\frac{1-t}{2}} \rho^{t/2} \exp\left(-\frac{t}{6} - \frac{1}{12}\right).$$

If $\min\{s, n-s\} < 2/\rho$, there are at most $(s+1)^{t-1}$ feasible count vectors in \mathcal{S} (compositions of s into t parts), so every pmf value is at least $(s+1)^{1-t}$; in particular $H(\mathbf{c}) \geq (s+1)^{1-t}$, which suffices for our purposes in the small- s regime. \square

We thus get the direct lower bound

$$\Pr_{\text{WOR}}(A) = \sum_{\mathbf{c} \in \mathcal{C}} H(\mathbf{c}) \geq H(\mathbf{c}^*) \geq Z'_t s^{-\frac{t-1}{2}}. \quad (\dagger)$$

Let $\delta' > 0$ be a target failure probability for Theorem 4.3 under WOR sampling. Applying Theorem 4.3 with failure probability δ' , if

$$|S| = \Omega\left(\frac{1}{\varepsilon^2} \log \frac{kn}{\delta'}\right).$$

then a WOR-uniform sample S of size $|S|$ is bad with probability at most $\Pr_{\text{WOR}}(B) \leq \delta'$.

Set $\delta' = \delta \cdot \Pr(A)$ and use (\dagger) to bound $\Pr(A)$ from below. This yields

$$|S| = \Omega\left(\frac{1}{\varepsilon^2} \left(\log \frac{kn}{\delta} + t \log |S|\right)\right).$$

Equivalently, by a standard bootstrapping argument, this is implied by

$$|S| = \Omega\left(\frac{1}{\varepsilon^2} \left[\log \frac{n}{\delta} + t \log \left(\frac{t}{\varepsilon^2} \log \frac{n}{\delta}\right)\right]\right),$$

where we used $k \leq n$.

With this choice, $\Pr(B) \leq \delta'$. Therefore,

$$\Pr(B | A) = \frac{\Pr(B \cap A)}{\Pr(A)} \leq \frac{\Pr(B)}{\Pr(A)} \leq \frac{\delta'}{\Pr(A)} = \delta.$$

This establishes the theorem: conditional on S being stratified (equivalently, sampling uniformly without replacement among stratified samples), a PRF slate for S is up to one 2-approximate PRF for \mathcal{N} with probability at most δ of failure. \square

E.3. Dimension Dependent Sampling Bounds independent of m and k

Lemma E.5 (Uniform preservation of witness sets). *Let \mathcal{M} be a finite population of size m , and let $\mathcal{W} \subseteq 2^{\mathcal{M}}$ be a family of witness sets with VC dimension at most h . Let $\mathcal{N} \subseteq \mathcal{M}$ be a uniformly random sample of n agents drawn without replacement. If $n = O(\varepsilon^{-2}(h \log(1/\varepsilon) + \log(1/\delta)))$, then with probability at least $1 - \delta$, every $W \in \mathcal{W}$ satisfies*

$$\left| \frac{|W \cap \mathcal{N}|}{|\mathcal{N}|} - \frac{|W|}{|\mathcal{M}|} \right| \leq \varepsilon.$$

Proof. For $W \in \mathcal{W}$, let $p(W) = |W|/|\mathcal{M}|$ and $\hat{p}(W) = |W \cap \mathcal{N}|/|\mathcal{N}|$. By the VC inequality, for a sample of size n ,

$$\Pr\left(\sup_{W \in \mathcal{W}} |\hat{p}(W) - p(W)| > \varepsilon\right) \leq 8\Pi_{\mathcal{W}}(2n) \exp(-n\varepsilon^2/32).$$

Since $\text{VCdim}(\mathcal{W}) \leq h$, Sauer–Shelah gives $\Pi_{\mathcal{W}}(2n) \leq (2en/h)^h$. Hence the failure probability is at most

$$8(2en/h)^h \exp(-n\varepsilon^2/32).$$

Thus, it suffices that

$$n \geq \frac{32}{\varepsilon^2} \left(h \log \frac{2en}{h} + \log \frac{8}{\delta} \right).$$

Claim E.6. Let $L := \log(8/\delta)$, and define $f(m) := \frac{32}{\varepsilon^2} (h \log \frac{2em}{h} + L)$. There exists an absolute constant C such that if $m_0 := C(h \log(1/\varepsilon) + L)/\varepsilon^2$, then every $m \geq m_0$ satisfies $m \geq f(m)$.

Proof. First, f is increasing and concave with derivative $f'(m) = \frac{64}{\varepsilon^2} \frac{v}{m}$. Hence for all $m \geq m_* := \frac{64}{\varepsilon^2} v$ we have $f'(m) \leq 1$. Choose $m_0 := C(v \log(1/\varepsilon) + L)/\varepsilon^2$ with C large enough so that $m_0 \geq m_*$. We upper bound the logarithm at m_0 by

$$\log \frac{2em_0}{v} \leq \log \frac{m_0}{v} + \log(2e) = \log(2Ce) + 2 \log(1/\varepsilon) + \log(\log(1/\varepsilon) + L/v)$$

Using that $\log(a + b) \leq \log(a) + b/a$ for $a, b > 0$ and setting $M = \log(1/\varepsilon)$,

$$\log(2Ce) + 2 \log(1/\varepsilon) + \log(M + L/v) \leq \log(2Ce) + 3M + \frac{L}{vM} \quad (5)$$

since $\log(1/\varepsilon) > 0$ ($\varepsilon < 1$), $\log(1/\delta) > 0$ ($\delta < 1$). Therefore

$$\begin{aligned} f(m_0) &\leq \frac{32}{\varepsilon^2} \left(v \left[\frac{L}{vM} + 3M + \log(2Ce) \right] + L \right) \\ &= \frac{32}{\varepsilon^2} \left(\frac{L}{M} + 3vM + v \log(2Ce) + L \right) \\ &\leq \frac{32}{\varepsilon^2} \left(3vM + v \log(2Ce) + 2L \right). \end{aligned}$$

Comparing with $m_0 = \frac{C}{\varepsilon^2} (vM + L)$, choosing $C \geq 64$ large enough ensures $f(m_0) \leq m_0$ for all $v \geq 1$ and $L \geq 0$.

Finally, for any $m \geq m_0$, concavity with $f'(m_0) \leq 1$ gives the supporting-line bound $f(m) \leq f(m_0) + f'(m_0)(m - m_0) \leq f(m_0) + (m - m_0) \leq m$, so $m \geq f(m)$ for all $m \geq m_0$, as claimed. \square

The claim implies that $n = O(\varepsilon^{-2}(h \log(1/\varepsilon) + \log(1/\delta)))$ is sufficient for the failure probability to be at most δ . This proves the lemma. \square

Given a finite population $\mathcal{P} \subseteq \mathcal{X}$ and a range family $\mathcal{R} \subseteq 2^{\mathcal{X}}$, write

$$\mathcal{F}_{\mathcal{R}}(\mathcal{P}) := \{R \cap \mathcal{P} : R \in \mathcal{R}\}$$

for the subsets of the population induced by \mathcal{R} .

We state the next corollary for Euclidean balls, for which the induced range family has VC dimension $d + 1$. The same argument applies to any range family \mathcal{R} with VC dimension h , replacing d in the sample bound by h .

Corollary E.7 (No-loss transfer for ball-induced groups). *Let $\mathcal{M} \subset \mathbb{R}^d$ be a population of size m , and let $\mathcal{N} \subseteq \mathcal{M}$ be a uniform sample of size*

$$n = O\left(\varepsilon^{-2} \left(d \log \frac{1}{\varepsilon} + \log \frac{1}{\delta} \right)\right).$$

Let \mathcal{R} be the range family of Euclidean balls in \mathbb{R}^d , and set $\mathcal{F} := \mathcal{F}_{\mathcal{R}}(\mathcal{P})$. If W^ is a PRF slate with respect to \mathcal{N} , then with probability at least $1 - \delta$, for every $S \in \mathcal{F}$ and every $\ell \in [k]$, if $|S|/M \geq \ell/k + \varepsilon$, then*

$$\left| \{c \in W^* \mid \min_{i \in S} \Delta(c, v_i) \leq D(S)\} \right| \geq \ell.$$

Proof. Let \mathcal{R} be the range family of Euclidean balls in \mathbb{R}^d , and let $\mathcal{F} := \{R \cap \mathcal{P} : R \in \mathcal{R}\}$ be the induced family of ball-induced subsets of the population. For $S \in \mathcal{F}$, define $p(S) := |S|/M$ and $\hat{p}_{\mathcal{N}}(S) := |S \cap \mathcal{N}|/n$. Since \mathcal{F} is induced by Euclidean balls and $\text{VCdim}(\mathcal{R}) = d + 1$, the same VC bound applies to \mathcal{F} . Thus, for

$$n = O\left(\frac{1}{\varepsilon^2} \left(d \log \frac{1}{\varepsilon} + \log \frac{1}{\delta} \right)\right),$$

we have, with probability at least $1 - \delta$,

$$\sup_{S \in \mathcal{F}} |\hat{p}_{\mathcal{N}}(S) - p(S)| \leq \varepsilon.$$

On this event, if $S \in \mathcal{F}$ satisfies $p(S) \geq \ell/k + \varepsilon$, then $\hat{p}_{\mathcal{N}}(S) \geq \ell/k$. Since W^* is a PRF slate with respect to \mathcal{N} , applying PRF to $S \cap \mathcal{N}$ gives at least ℓ members $c \in W^*$ such that $\Delta(c, S \cap \mathcal{N}) \leq \Delta(S \cap \mathcal{N})$. Because $S \cap \mathcal{N} \subseteq S$, we have $\Delta(c, S) \leq \Delta(c, S \cap \mathcal{N})$ and $\Delta(S \cap \mathcal{N}) \leq \Delta(S)$ for each such c . Hence $\Delta(c, S) \leq \Delta(S)$, proving the claim. \square

Theorem E.8 (Population transfer for arbitrary cohesive groups). *Let $\mathcal{P} \subset \mathbb{R}^d$ be a population of size M , and let k be the slate size. Let $\delta > 0$ and $\varepsilon > 0$, and let $\mathcal{N} \subseteq \mathcal{P}$ be sampled uniformly at random without replacement of size*

$$n = \Omega\left(\frac{1}{\varepsilon^2} \left(d \log \frac{1}{\varepsilon} + \log \frac{1}{\delta} \right)\right).$$

If W^* is a PRF slate with respect to \mathcal{N} , then with probability at least $1 - \delta$, the following holds: for every $\ell \in [k]$, every $S \subseteq \mathcal{P}$, and every $t > 0$ such that $|S| \geq (\ell/k + \varepsilon)M$ and S is t -cohesive,

$$\left| \left\{ c \in W^* \mid \exists i \in S, \Delta(i, c) \leq (1 + \sqrt{2})t \sqrt{\frac{d}{d+1}} \right\} \right| \geq \ell.$$

Proof. By the assumed lower bound on $|\mathcal{N}|$, Corollary E.7 applies to the sampled population \mathcal{N} . Hence, with probability at least $1 - \delta$, every ball-induced subset $S' \subseteq \mathcal{P}$ with $|S'|/M \geq \ell/k + \varepsilon$ satisfies

$$|\{c \in W^* : \Delta(c, S') \leq \Delta(S')\}| \geq \ell.$$

We condition on this event for the remainder of the proof.

Now consider a group S as in the statement of the Theorem. Let $B(c, R)$ be a minimum-radius Euclidean ball containing S . We now use the following standard form of Jung's theorem: any subset of \mathbb{R}^d with diameter at most t is contained in a Euclidean ball of radius at most $t\sqrt{d/(2(d+1))}$. Since $D(S) \leq t$, Jung's theorem gives

$$R \leq t \sqrt{\frac{d}{2(d+1)}}.$$

Let $S_B := B(c, R) \cap \mathcal{P}$. Since $S \subseteq S_B$ and $|S| \geq (\ell/k + \varepsilon)M$, we have $\frac{|S_B|}{M} \geq \frac{\ell}{k} + \varepsilon$. Moreover, $S_B \in \mathcal{F}$, since it is induced by the Euclidean ball $B(c, R)$. Therefore, by Corollary E.7, there are at least ℓ members $x \in W^*$ such that $\Delta(x, S_B) \leq \Delta(S_B)$. Finally, because $S_B \subseteq B(c, R)$, any two points in S_B are at distance at most $2R$. Hence $\Delta(S_B) \leq 2R$, and so for each such x ,

$$\Delta(x, S_B) \leq 2R.$$

Equivalently, for each such x , there exists some $z \in S_B = B(c, R) \cap \mathcal{P}$ with $\Delta(x, z) \leq 2R$.

It remains to relate z back to the original group S .

Let $A := S \cap \partial B(c, R)$ be the set of points of S on the boundary of the minimum enclosing ball. Since $B(c, R)$ is minimum-radius, its center c lies in $\text{conv}(A)$. Hence there exist points $a_i \in A$ and coefficients $\lambda_i \geq 0$, with $\sum_i \lambda_i = 1$, such that $\sum_i \lambda_i(a_i - c) = 0$.

Fix $z \in S_B$, and set $u := z - c$. Since $z \in B(c, R)$, we have $\Delta(z, c) \leq R$. If $\langle u, a_i - c \rangle < 0$ for all $a_i \in A$, then taking the inner product with the convex combination above gives

$$0 = \left\langle u, \sum_i \lambda_i(a_i - c) \right\rangle = \sum_i \lambda_i \langle u, a_i - c \rangle < 0,$$

a contradiction. Therefore, for some $a \in A \subseteq S$, we have $\langle z - c, a - c \rangle \geq 0$, i.e., the angle between $z - c$ and $a - c$ is at most 90° .

By the cosine law,

$$\Delta(z, a)^2 = \Delta(z, c)^2 + \Delta(a, c)^2 - 2\langle z - c, a - c \rangle \leq R^2 + R^2 = 2R^2,$$

where we used $\Delta(z, c) \leq R$, $\Delta(a, c) = R$, and $\langle z - c, a - c \rangle \geq 0$. Thus $\Delta(z, a) \leq \sqrt{2}R$.

Combining the two bounds, for each of the ℓ selected points x above there exists $a \in S$ such that

$$\Delta(x, a) \leq \Delta(x, z) + \Delta(z, a) \leq 2R + \sqrt{2}R = (2 + \sqrt{2})R.$$

Using Jung's bound on R , we obtain

$$\Delta(x, a) \leq (2 + \sqrt{2})y \sqrt{\frac{d}{2(d+1)}} = (1 + \sqrt{2})y \sqrt{\frac{d}{d+1}}.$$

Thus at least ℓ members of W^* are within the claimed distance of some member of S . \square

1485 F. Additional Results

1486 F.1. No Overrepresentation Property

1487 We discuss an axiom that demands no part of the viewpoint embedding space is *overrepresented*.

1488 **Definition F.1** (No Over-Representation Property (Jerrett & Anshelevich, 2025)). We say that X satisfies the *no over-*
 1489 *representation property (NORP)* if, for every integer $\ell \in \{2, \dots, k\}$ and every subset $T \subseteq X$ with $|T| = \ell$, we have

$$1490 \frac{|\{i \in \mathcal{N} \mid \Delta(u_i, T) \leq \Delta(T)\}|}{n} > \frac{\ell - 1}{k}.$$

1491 In words, NORP dictates that whenever a slate X contains a set of responses T that represent similar viewpoints (i.e., has a
 1492 low diameter $\Delta(T)$), there are sufficiently many agents in the population that are well-represented by these responses (i.e.,
 1493 they are close in the viewpoint embedding space) to justify the inclusion of these responses in X .

1494 Just as with PRF, k -means can fail NORP. To see this, consider again Example A.1. Two of the four responses output
 1495 by k -means are arbitrarily close to $x = 1$, and hence highly similar to one another, even though 98 percent of agents are
 1496 far from these responses and therefore dislike them. Thus the region around $x = 1$ is overrepresented, violating NORP
 1497 (Definition F.1), including for $\ell = 2$.

1498 Jerrett & Anshelevich (2025) show that a modification of SEAR from Aziz et al. (2024) satisfies NORP while preserving
 1499 PRF. This NORP- and PRF-satisfying modification is the version of SEAR presented in this paper; our streamlined amortized
 1500 implementation idea was directly compatible with this version.

1501 Both of the axioms PRF and NORP are properties of a slate relative to the underlying agent population and should be
 1502 understood as necessary representation requirements, not as a complete characterization of proportional representation.

1503 For example, if a slate fails PRF, especially by a large margin, then it is natural to say that some sufficiently large and
 1504 cohesive group of agents is underrepresented. Similarly, if a slate fails NORP by a large margin, then some narrow region of
 1505 realized viewpoint space is overrepresented relative to the density of nearby realized viewpoints in the population.

1506 F.2. Selection–Generation Theorem for NORP

1507 **Theorem F.2** (Selection–Generation Theorem Commutation). Let $\{\phi_i(p)\}_{i \in \mathcal{N}}$ be γ -proxy viewpoint embeddings.

1508 For every $T \subseteq Y$, $|T| = \ell$ with $D(T) = t$, the size of the set of agents S such that for every $i \in S$

$$1509 \Delta(\psi(o), \psi(o_i^*)) \leq 2\gamma + t,$$

1510 for some $o \in T$ is greater than $\ell^{-1/k} \cdot n$.

1511 *Proof.* Select-then-Generate returns a slate of responses $Y = \{A_i(p) \mid i \in R\}$. Each $o \in y$ is generated by using a
 1512 simulation query from agent i 's simulation on prompt p and this $i \in R$ corresponds to a viewpoint in X selected by SEAR
 1513 run when on the proxy embeddings $\{\phi_i(p)\}$. We now show that NORP over the proxy viewpoint embeddings, the condition
 1514 implied by SEAR, implies approximate approximate NORP with respect to the true embeddings $\{\psi(A_i(p))\}_{i \in \mathcal{N}}$, where
 1515 the approximation is as stated in the theorem.

1516 Let Z be a subset of R of size ℓ , and suppose that Z has diameter at most t , i.e. $\max_{i,j} \Delta(\phi_i(p), \phi_j(p)) \leq t$. Let $T \subseteq Y$ be
 1517 the corresponding set of responses. By definition of γ -proxy embeddings, we have that for any $i \in \mathcal{N}$

$$1518 \Delta(\phi_i(p), \psi(A_i(p))) \leq \gamma.$$

1519 If in addition for some $j \in Z$ we have

$$1520 \Delta(\phi_i(p), \phi_j(p)) \leq t,$$

1521 then by the triangle inequality it follows that

1522 $\Delta(\phi_j(p), \psi(A_i(p))) \leq t + \gamma$. The NORP guarantee applied to the proxy embeddings thus translates that $|\{i \in \mathcal{N} : \exists j \in$

1540 Z s.t. $|\Delta(\phi_j(p), \phi_i(p)) \leq t\}| > \frac{\ell-1}{k}n$

$$\begin{aligned}
 1541 & \\
 1542 & \frac{\ell-1}{k}n < |\{i \in \mathcal{N} : \exists j \in Z \text{ s.t. } \Delta(\phi_i(p), \phi_j(p)) \leq t\}| \\
 1543 & \\
 1544 & \leq |\{i \in \mathcal{N} : \exists j \in Z \text{ s.t. } \Delta(\psi(A_i(p)), \phi_j(p)) \leq \gamma + t\}| \\
 1545 & \leq |\{i \in \mathcal{N} : \exists j \in Z \text{ s.t. } \Delta(\psi(A_i(p)), \psi(A_j(p))) \leq 2\gamma + t\}| \\
 1546 & = |\{i \in \mathcal{N} : \exists o \in T \text{ s.t. } \Delta(\psi(o_i^*), \psi(o)) \leq 2\gamma + t\}|. \\
 1547 &
 \end{aligned}$$

1548 as claimed. □

1550 F.3. Learning Viewpoint Embeddings via Reward Model

1552 We do not assume *a priori* access to the viewpoint embedding function ψ , nor that generic embedding models necessarily
 1553 capture viewpoint similarity. When off-the-shelf embeddings are insufficient, the relevant geometry can be learned from
 1554 viewpoint-relevant supervision, such as comparison data. A natural way to model human choice behavior is via a reward
 1555 function that assigns higher scores to responses that an agent prefers for a given prompt. In standard RLHF pipelines, one
 1556 typically trains a single reward model intended to capture aggregate human judgment, and disagreement across annotators is
 1557 often treated as noise. In our setting, however, disagreements between agents is precisely the signal we wish to preserve.

1559 Learning a fully separate reward model for each agent is typically infeasible. We therefore turn to personalized reward
 1560 models, which are inspired by collaborative filtering and posit that heterogeneous preferences admit a shared low-dimensional
 1561 structure.

1562 A factorized personalized reward model provides one way to turn such supervision into a viewpoint embedding space. For
 1563 the viewpoint setting, we introduce a factorized architecture. For each agent $i \in \mathcal{N}$, let $r_i(p, o)$ denote the reward that agent
 1564 i assigns to output $o \in \mathcal{V}^*$ in response to prompt $p \in \mathcal{V}^*$. We say that agent i has a *factorized viewpoint reward model* if

$$1565 \quad r_i(p, o) = \psi(p, o)^\top \phi_i(p),$$

1568 where $\psi : \mathcal{V}^* \times \mathcal{V}^* \rightarrow \mathbb{S}^{d-1}$ is a shared embedding map for prompt-response pairs, and $\phi_i : \mathcal{V}^* \rightarrow \mathbb{S}^{d-1}$ is agent i 's
 1569 prompt-conditioned viewpoint embedding. Only the final interaction is assumed to be linear; both ψ and the maps ϕ_i may
 1570 themselves be arbitrarily rich function classes. Note that L_2 normalization of embeddings is standard in the literature.

1572 The distinction between the two terms is important. The shared map $\psi(p, o)$ captures properties of a candidate output *in*
 1573 *the context of the prompt*. Prompt dependence in ψ is therefore essential whenever the same response may be appropriate
 1574 for one prompt and inappropriate for another. By contrast, $\phi_i(p)$ captures what agent i values *for that prompt*: it can be
 1575 interpreted as a prompt-conditioned weighting over latent attributes. Thus prompt dependence in ϕ_i captures the possibility
 1576 that the same agent may care about different latent dimensions in different contexts.

1577 This architecture becomes a viewpoint model under an additional ideal-point assumption. Suppose that, for each agent i and
 1578 prompt p , there exists a response o_i^* maximizing agent i 's reward, and thus that the corresponding embedding realizes the
 1579 agent's ideal point:

$$1580 \quad \phi_i(p) = \psi(p, o_i^*).$$

1582 Under this assumption, the shared map ψ induces a viewpoint embedding space, and the point $\psi(p, o_i^*)$ can be interpreted as
 1583 agent i 's viewpoint for prompt p . The reason is immediate from the geometry of the unit sphere. Thus, if the personalized
 1584 reward models are learned accurately enough, the shared embedding map ψ provides exactly the kind of viewpoint
 1585 embedding space needed for our representation framework.

1587 Note that a generic personalized reward model may learn a shared latent representation of responses, but this does not by
 1588 itself yield a viewpoint embedding space: preferences may still be highly non-spatial in that representation. Our architecture
 1589 is more restrictive. By requiring both response embeddings and agent viewpoint embeddings to lie on the unit sphere and
 1590 interact through an inner product, we obtain a geometry in which each agent's reward is monotone in distance to an ideal
 1591 point. Thus the learned latent space is not merely a shared embedding space, but a viewpoint embedding space in the sense
 1592 required for our representation framework.

1593 These personalized reward models admit a natural interpretation in the context of Section 4.

Latent Reward Perspective

A factorized personalized reward model

$$r_i(p, o) = \psi(p, o)^\top \phi_i(p)$$

learns a shared prompt-response embedding $\psi(p, o)$ and a prompt-specific agent embedding $\phi_i(p)$. Then, if we can generate or approximately optimize responses o_i^* such that $r_i(p, o_i^*) \geq \gamma$, then the learned embeddings $\{\phi_i(p)\}_{i \in \mathcal{N}}$ are γ -proxy viewpoint embeddings with respect to the jointly learned embedding space ψ .

We can therefore run SEAR on the learned predicted embeddings and apply the approximate representation guarantee of Theorem 4.2. Alternatively, we can prove a reward version of the Selection-Generation Commutation Theorem that treats reward models as first class-citizens: First, we state a basic technical lemma.

Lemma F.3. *Let $a, b, c \in \mathbb{S}^{d-1}$ be unit vectors. If $a \cdot b \geq t_1$ and $b \cdot c \geq t_2$, then*

$$a \cdot c \geq t_1 t_2 - \sqrt{(1 - t_1^2)(1 - t_2^2)}.$$

Proof. Decompose a and c into parts parallel and orthogonal to b :

$$a = (a \cdot b)b + a_\perp, \quad c = (c \cdot b)b + c_\perp,$$

with $a_\perp \cdot b = c_\perp \cdot b = 0$ and $\|a_\perp\| = \sqrt{1 - (a \cdot b)^2}$, $\|c_\perp\| = \sqrt{1 - (c \cdot b)^2}$. Then

$$\begin{aligned} a \cdot c &= (a \cdot b)(c \cdot b) + a_\perp \cdot c_\perp \geq t_1 t_2 - \|a_\perp\| \|c_\perp\| \\ &\geq t_1 t_2 - \sqrt{(1 - t_1^2)(1 - t_2^2)}, \end{aligned}$$

using $a \cdot b \geq t_1$, $c \cdot b \geq t_2$, and Cauchy–Schwarz with a_\perp antiparallel to c_\perp for the worst case. \square

Theorem F.4 (Selection–Generation, Reward Model Version). *For each agent i , let A_i be any black-box simulation model with the property that, on prompt p , it returns a response whose embedding approximately maximizes the factorized reward $r_i(p, o)$ so that*

$$r_i(p, A_i(p)) \geq \gamma$$

for all $i \in \mathcal{N}$. The slate Y returned by Select-then-Generate on $\{\phi_i(p)\}_{i \in \mathcal{N}}$ has the following guarantees w.r.t. to the agent reward models.

1. *For every ℓ -large and t -cohesive group of agents $S \subseteq \mathcal{N}$ there exist at least ℓ outputs $o \in Y$ such that for each such o ,*

$$r_i(p, o) \geq t\gamma - \sqrt{(1 - t^2)(1 - \gamma^2)},$$

for some $i \in S$.

2. *For every $T \subseteq Y$, $|T| = \ell$ with pairwise cosine similarity at least t , i.e. there exists $t > 0$ such that $\psi(p, o_1)^\top \psi(p, o_2) \geq t$ for all $o_1, o_2 \in T$, the set of agents S such that for every $i \in S$*

$$r_i(p, o) \geq t\gamma - \sqrt{(1 - t^2)(1 - \gamma^2)},$$

for some $o \in T$ is greater than $\ell^{-1/k} \cdot n$.

Proof. Since the simulation A_i approximately maximizes each agent i 's reward function, for any prompt p and agent i , the model produces a response

$$\begin{aligned} A_i(p) &\in \{o \in \mathcal{V}^* \mid r_i(p, o) \geq \gamma\} \\ &= \{o \in \mathcal{V}^* \mid \psi(p, o)^\top \phi_i(p) \geq \gamma\}. \end{aligned}$$

Now, consider the SEAR algorithm (Appendix A.3), which, given agent embeddings $\{\phi_i(p)\}_{i \in \mathcal{N}}$, selects a slate $R \subseteq \mathcal{N}$ of k agents such that their viewpoint embeddings $\{\phi_i(p)\}_{i \in R}$ satisfy PRF in embedding space. Select-Then-Generate returns a slate of responses $Y = \{A_i(p) : i \in R\}$ generated by k simulation queries to each of the agents in R . Let $g(t, \gamma) = t\gamma - \sqrt{(1-t^2)(1-\gamma^2)}$. We will show that PRF (Definition 3.1) and NORP (Definition F.1) over the embeddings $\phi_i(p)$, the condition implied by SEAR, implies approximate reward level guarantees for ℓ -large and t -cohesive groups, as stated in the theorem.

By assumption on A_x , we have that for any $x \in R$

$$r_x(p, A_x(p)) = \phi_i(p)^\top \psi(p, A_x(p)) \geq \gamma.$$

If in addition for some $i \in S$ we have

$$\phi_i(p)^\top \phi_x(p) \geq t,$$

then by Lemma F.3 it follows that $\phi_i(p)^\top \psi(p, A_x(p)) \geq g(t, \gamma)$.

PRF Therefore, the PRF representation guarantee that $|\{x \in X : \exists i \in S \text{ s.t. } \phi_i(p)^\top \phi_x(p) \geq t\}| \geq \ell$ translates to

$$\begin{aligned} \ell &\leq |\{x \in X : \exists i \in S \text{ s.t. } \phi_i(p)^\top \phi_x(p) \geq t\}| \\ &\leq |\{x \in X : \exists i \in S \text{ s.t. } \phi_i(p)^\top \psi(p, A_i(p)) \geq g(t, \gamma)\}| \\ &= |\{x \in X : \exists i \in S \text{ s.t. } r_i(p, A_i(p)) \geq g(t, \gamma)\}| \\ &= |\{o \in Y : \exists i \in S \text{ s.t. } r_i(p, o) \geq g(t, \gamma)\}|. \end{aligned}$$

We conclude that the slate Y produced by generating responses from the k selected agent embeddings approximately satisfies PRF over the space of responses:

For every ℓ -large and t -cohesive group of agents $S \subseteq \mathcal{N}$ there exist ℓ outputs $o \in Y$ such that for each such o ,

$$r_i(p, o) \geq g(t, \gamma).$$

for some $i \in S$.

NORP Fix $Z \subseteq X$ and let $t = \min_{i,j \in Z} \phi_i(p)^\top \phi_j(p)$. Let $T = \{A_x(p) \mid x \in Z\}$. The NORP guarantee that $|\{i \in \mathcal{N} : \exists j \in Z \text{ s.t. } \phi_j(p)^\top \phi_i(p) \geq t\}| > \frac{\ell-1}{k}n$ translates to

$$\begin{aligned} \frac{\ell-1}{k}n &< |\{i \in \mathcal{N} : \exists j \in Z \text{ s.t. } \phi_j(p)^\top \phi_i(p) \geq t\}| \\ &= |\{i \in \mathcal{N} : \exists j \in Z \text{ s.t. } \psi(p, A_j(p))^\top \phi_i(p) \geq g(t, \gamma)\}| \\ &= |\{i \in \mathcal{N} : \exists o \in T \text{ s.t. } \psi(p, o)^\top \phi_i(p) \geq g(t, \gamma)\}| \\ &= |\{i \in \mathcal{N} : \exists o \in T \text{ s.t. } r_i(p, o) \geq g(t, \gamma)\}|. \end{aligned}$$

□

G. Experimental Details

G.1. Topic Generation Prompt

The following prompt was given to Claude Opus 4.6¹⁶ to generate ten diverse topics for evaluation:

Here are four questions that a group of participants was asked about:
 What are your impressions of the campus protests?
 What are your impressions of how campus administrators handled the protests?
 What are your thoughts on the way university campus administrators should approach the issue of Israel/Gaza demonstrations?
 What should guide university campus administrators handling of protests?
 Given a person's answer to these questions, what are ten other diverse topics would be predictable*?
 Return only a JSONL object where each row is formatted as {"topic": topic}. The topics you generated will be given to an LLM to generate a variety of questions on that topic, so give enough information about what each topic represents in the JSONL.

¹⁶Claude Opus 4.6 (Anthropic, 2026) is released under a commercial license by Anthropic.

1705 **G.2. Prompt Generation Prompt**

1706 For each of the ten generated topics, the following prompt template was given to Claude Opus 4.6 to generate 100 diverse
1707 opinion questions:
1708

1709 Generate 100 diverse and open-ended questions asking people for their opinions on the following topic: {topic}. Choose a
1710 mix of contentious questions that people are likely divided on and non-contentious questions that most people would be in
1711 agreement on. You should make each question stand-alone, so that a person can understand what you are referring to without
1712 additional context. Your output format should be a jsonl object where each line is formatted as {"prompt_id": prompt_id,
1713 "prompt": prompt}

1714
1715 **G.3. Generated Topics**

1716 The following ten topics were generated by Claude Opus 4.6:
1717

- 1718 1. Views on free speech and the First Amendment, including where the line should be drawn between protected expression
1719 and harassment or disruption
- 1720 2. Opinions on the Israel-Palestine conflict, including views on Israeli military operations in Gaza and Palestinian
1721 statehood
- 1722 3. Attitudes toward higher education institutions, including trust in university leadership, the purpose of universities, and
1723 whether colleges are too liberal or too conservative
- 1724 4. Views on student activism and civil disobedience throughout history, including whether disruptive protest is an effective
1725 or justified means of social change
- 1726 5. Opinions on police use of force in crowd control and protest situations, including whether law enforcement intervention
1727 on campuses is appropriate
- 1728 6. Attitudes toward U.S. foreign policy and military aid, particularly regarding funding and arms sales to allied nations in
1729 conflict zones
- 1730 7. Views on antisemitism and Islamophobia, including how to define hate speech, whether certain political criticisms
1731 cross into bigotry, and how institutions should address discrimination
- 1732 8. Opinions on generational differences in political engagement, including whether younger generations are more or less
1733 informed, idealistic, or effective in their activism compared to older generations
- 1734 9. Attitudes toward media coverage and bias, including whether mainstream and social media fairly represent protest
1735 movements, campus issues, and the Israel-Gaza conflict
- 1736 10. Views on corporate and institutional divestment as a political tool, including whether universities, pension funds, and
1737 businesses should divest from industries or nations involved in controversial practices

1738
1739 **G.4. Example Prompts**

1740 Below we list five example prompts from each of the ten generated topics.
1741

1742 **Topic 1: Free speech and the First Amendment.**

- 1743 1. Do you believe the First Amendment's protection of free speech is one of the most important rights guaranteed by the
1744 U.S. Constitution? Why or why not?
 - 1745 2. Should social media companies have the right to ban users for speech that is legal but violates their terms of service, or
1746 does this undermine the spirit of free speech?
 - 1747 3. Where do you personally draw the line between expressing a controversial opinion and engaging in harassment?
- 1748
1749
1750
1751
1752
1753
1754
1755
1756
1757
1758
1759

Social Choice Foundations for Simulation-Augmented Generation

- 1760 4. Do you think hate speech should be legally protected under the First Amendment, even if it targets people based on
1761 race, gender, religion, or sexual orientation?
1762
- 1763 5. Should public universities be allowed to disinvite controversial speakers if students protest against their presence on
1764 campus?
1765

Topic 2: Israel-Palestine conflict.

- 1766
1767
1768 1. Do you believe a two-state solution is still a viable path to peace between Israel and Palestine? Why or why not?
1769
1770 2. In your opinion, should the international community do more to protect civilian lives in Gaza during Israeli military
1771 operations?
1772
1773 3. Do you think both Israeli and Palestinian children deserve to grow up in safety and security, regardless of the political
1774 situation?
1775
1776 4. What is your opinion on Israel's use of airstrikes in densely populated areas of Gaza?
1777
1778 5. Do you believe Palestine should be recognized as an independent state by all countries in the United Nations?
1779
1780

Topic 3: Higher education institutions.

- 1781
1782
1783 1. Do you believe that the primary purpose of a university education is to prepare students for a career, or to provide a
1784 broad intellectual foundation? Why?
1785
1786 2. How much trust do you have in university presidents and chancellors to make decisions that are in the best interest of
1787 students?
1788
1789 3. Do you think colleges and universities in the United States are generally too liberal in their political orientation? Why
1790 or why not?
1791
1792 4. Do you believe a four-year college degree is still worth the financial investment it requires today?
1793
1794 5. Should university administrators be required to have recent classroom teaching experience in order to lead an academic
1795 institution?
1796
1797

Topic 4: Student activism and civil disobedience.

- 1798
1799
1800 1. Do you think student activists throughout history have played an important role in bringing about positive social
1801 change?
1802
1803 2. In your opinion, is civil disobedience ever morally justified, even when it involves breaking the law?
1804
1805 3. Do you believe the student-led sit-ins during the U.S. Civil Rights Movement of the 1960s were an effective strategy
1806 for ending racial segregation?
1807
1808 4. Should universities punish students who participate in protests that disrupt campus operations, such as blocking
1809 buildings or occupying administrative offices?
1810
1811 5. Do you think there is a meaningful difference between peaceful protest and civil disobedience, and should society treat
1812 them differently?
1813
1814

1815 **Topic 5: Police use of force.**

- 1816
- 1817 1. Do you think police should ever use tear gas to disperse peaceful protesters, or should tear gas only be reserved for
- 1818 situations involving violence?
- 1819
- 1820 2. In your opinion, is it appropriate for universities to call in outside law enforcement to handle student protests on
- 1821 campus?
- 1822
- 1823 3. Do you believe that police officers generally use an appropriate level of force when managing large protest crowds, or
- 1824 do they tend to escalate situations unnecessarily?
- 1825
- 1826 4. Should rubber bullets be considered a legitimate crowd control tool during protests, given the documented injuries they
- 1827 can cause?
- 1828
- 1829 5. Do you think there should be an independent civilian review board that evaluates every instance of police use of force
- 1828 during protests?

1830 **Topic 6: U.S. foreign policy and military aid.**

- 1831
- 1832 1. Do you think the United States has a moral obligation to provide military aid to its allied nations when they are involved
- 1833 in armed conflicts?
- 1834
- 1835 2. How do you feel about U.S. taxpayer dollars being used to fund arms sales to countries engaged in active war zones?
- 1836
- 1837 3. In your opinion, should the U.S. government prioritize domestic spending over foreign military aid to allied nations?
- 1838
- 1839 4. Do you believe that U.S. arms sales to allied nations in conflict zones ultimately make the world safer or more
- 1840 dangerous?
- 1841
- 1842 5. Should the United States place strict human rights conditions on military aid it provides to allied countries involved in
- 1841 conflicts?

1843 **Topic 7: Antisemitism and Islamophobia.**

- 1844
- 1845 1. Do you believe antisemitism is still a significant problem in modern society, or has it largely been addressed?
- 1846
- 1847 2. Do you believe Islamophobia is still a significant problem in modern society, or has it largely been addressed?
- 1848
- 1849 3. How would you personally define hate speech? Where do you draw the line between offensive speech and hate speech?
- 1850
- 1851 4. Do you think criticizing the policies of the Israeli government is inherently antisemitic, or can such criticism be
- 1852 legitimate political discourse?
- 1853
- 1854 5. Do you think criticizing Islamic religious practices or beliefs is inherently Islamophobic, or can such criticism be
- 1853 legitimate discourse?

1855 **Topic 8: Generational differences in political engagement.**

- 1856
- 1857 1. Do you think younger generations (Gen Z and Millennials) are more politically informed than older generations (Baby
- 1858 Boomers and Gen X), or do they just have access to more information without necessarily being better informed?
- 1859
- 1860 2. In your opinion, is social media activism (liking, sharing, posting about political causes) a meaningful form of political
- 1861 engagement, or is it mostly performative?
- 1862
- 1863 3. Do you believe that older generations tend to dismiss younger people's political opinions as naive or uninformed? If so,
- 1864 is that dismissal justified?
- 1865
- 1866 4. Do you think younger generations are more idealistic about politics compared to older generations, and is that idealism
- 1867 a strength or a weakness?
- 1868
- 1869 5. In your view, are older generations more effective at creating lasting political change because they work within
- 1868 established systems, or are younger generations more effective because they challenge those systems?

1870 **Topic 9: Media coverage and bias.**

- 1871
- 1872 1. Do you believe mainstream media outlets (such as CNN, Fox News, MSNBC, etc.) provide fair and balanced coverage
- 1873 of the Israel-Gaza conflict, or do they tend to favor one side? Please explain your reasoning.
- 1874
- 1875 2. In your opinion, has social media (platforms like Twitter/X, Instagram, TikTok, etc.) been a positive or negative force
- 1876 in shaping public understanding of the Israel-Gaza conflict?
- 1877
- 1878 3. Do you think pro-Palestinian campus protests at universities have been fairly represented by mainstream media, or has
- 1879 the coverage been skewed in a particular direction?
- 1880
- 1881 4. How important do you think it is for journalists to clearly distinguish between news reporting and opinion commentary
- 1882 when covering politically sensitive topics like the Israel-Gaza conflict?
- 1883
- 1884 5. Do you believe that social media algorithms contribute to political polarization by showing people only content that
- 1885 reinforces their existing views on issues like the Israel-Gaza conflict?

1886 **Topic 10: Corporate and institutional divestment.**

- 1887
- 1888 1. Do you think universities have a moral obligation to divest their endowments from fossil fuel companies, even if it
- 1889 means potentially lower financial returns for the institution?
- 1890
- 1891 2. Should public pension funds divest from companies that manufacture weapons used in armed conflicts abroad?
- 1892
- 1893 3. Do you believe corporate divestment campaigns are an effective way to pressure governments to change their policies,
- 1894 or are they mostly symbolic gestures?
- 1895
- 1896 4. Is it appropriate for university students to demand that their institution divest from specific industries or countries,
- 1897 given that students are temporary members of the university community?
- 1898
- 1899 5. Should pension funds prioritize maximizing financial returns for retirees over ethical considerations when making
- 1900 investment decisions?

1901 **G.5. Simulated Agent Prompt**

1902 Each simulated agent is implemented via a roleplay prompt to Llama-3.1-8B-Instruct (Grattafiori et al., 2024)¹⁷ with the

1903 following system and user prompts.

1904

1905 **System prompt.**

1906 You are roleplaying as the following person:

1907 *{agent description}*

1908 Generate a genuine response that reflects this person’s personal perspective. Do not hedge or speak in generalities — respond

1909 as this specific person would.

1910

1911

1912 The participant description is constructed from the participant’s demographic attributes in the Remesh survey (e.g., gender,

1913 age, political affiliation, religious affiliation, education level, urbanicity, household income).

1914 **User prompt.** The user prompt is composed of three optional sections, joined by blank lines:

1915

- 1916 1. **Previous responses** (if available):

1917 Here are questions you answered previously:

1918 Question: *{question text}*

1919 Response: *{participant’s response}*

1920 [...repeated for each prior question...]

1921 ¹⁷Llama-3.1-8B-Instruct (Grattafiori et al., 2024) is released under a commercial license from Meta.

1922

1923

1924

2. **Voting context** (if available):

Here are other people’s statements that you agreed with:
 Question: *{question text}*
 Statement: *{statement text}*
 [...repeated for each agreed statement...]
 Here are other people’s statements that you disagreed with:
 Question: *<question text>*
 Statement: *<statement text>*
 [...repeated for each disagreed statement...]

3. **Target question:**

Now answer this new question in 2-3 sentences.
 Question: *<target question text>*
 Response:

G.6. Viewpoint Summary Prompt

For each participant, we summarize their simulated agent’s responses on the train prompts into a concise viewpoint summary using Claude Opus 4.6. Responses are grouped by topic, and each topic is summarized separately using the following prompt:

System prompt.

You are a helpful assistant that summarizes a person’s viewpoints. Given a set of questions and the person’s responses on a particular topic, write a concise summary (<=3 sentences) capturing their key views, attitudes, and reasoning on this topic. Do not list individual responses — synthesize them into a coherent summary.

User prompt.

Topic: *{topic}*
 Here are the person’s responses:
 Q: *{question text}*
 A: *{viewpoint text}*
 [...repeated for each response on this topic...]
 Summarize this person’s views on this topic in three sentences or less.

The per-topic summaries are then concatenated into a single viewpoint summary per participant, formatted as ****<topic>****: **<summary>** for each topic. An example viewpoint summary for one simulated participant is shown below.

Example viewpoint summary.

Opinions on generational differences in political engagement: This person holds a balanced, nuanced view of generational differences in political engagement, seeing younger generations as bringing valuable energy, idealism, diversity, and digital savvy to activism while acknowledging that their reliance on social media makes them vulnerable to misinformation, oversimplification, and shallow engagement that may not translate into lasting policy change. They believe older generations offer important experience, mentorship, and effective traditional organizing methods, but can be dismissive of younger perspectives and may hold disproportionate political power through higher voter turnout. Overall, they advocate for intergenerational collaboration rather than “generational warfare,” viewing the strengths and weaknesses of each generation as complementary, and they see systemic barriers—not apathy—as the primary obstacle to younger people’s fuller participation in democracy.

Attitudes toward U.S. foreign policy and military aid: This person is deeply critical of U.S. military aid and arms sales policies, viewing them as fundamentally escalatory, ethically inconsistent with American values of peace and human rights, and often driven more by economic and political interests than genuine security concerns. They strongly advocate for shifting U.S. foreign policy toward diplomacy, conflict prevention, humanitarian aid, and non-lethal alternatives, while demanding greater congressional oversight, strict human rights conditions, transparency, and public input in military aid decisions. Though they acknowledge some nuance—supporting limited aid in cases of existential threats like Ukraine or Taiwan and recognizing the value of alliances—they broadly believe domestic spending should take priority and that the U.S. cannot credibly promote peace while being the world’s largest arms exporter.

1980 **Attitudes toward media coverage and bias:** This person believes mainstream media coverage of the Israel-Gaza conflict and
1981 related campus protests is significantly biased, often sensationalized, and tends to favor pro-Israel narratives while unfairly
1982 portraying pro-Palestinian protests as disruptive or antisemitic, leading to an erosion of public trust. They advocate for greater
1983 media transparency, nuanced reporting, and consuming multiple diverse sources, while expressing concern that corporate
1984 ownership, domestic political dynamics, and the 24-hour news cycle undermine journalistic objectivity. They see international
1985 media and campus student outlets as generally more balanced than American mainstream media, view social media as a
1986 mixed but largely negative force due to algorithmic echo chambers, and support industry self-regulation and disclosure of
commentator affiliations rather than government-imposed restrictions.

1987 **Views on corporate and institutional divestment:** This person views divestment as a legitimate and sometimes morally
1988 necessary tool for political expression and social change—particularly regarding fossil fuels, private prisons, weapons
1989 manufacturers, and companies with exploitative practices—while consistently emphasizing that it should be part of a broader
1990 strategy combining sanctions, policy reform, and public pressure rather than a standalone solution. They are generally
1991 supportive of ESG investing and institutional accountability but express nuanced concerns about selective targeting, the
1992 potential for moral posturing, the risk of politicizing institutions like pension funds, and the unintended consequences
1993 divestment can have on workers and industries’ ability to reform from within. They strongly oppose anti-BDS and anti-
1994 divestment laws as violations of free speech, believe final institutional decisions should rest with boards of trustees after
inclusive stakeholder input, and favor giving targeted companies an opportunity to reform before divestment is pursued.

1995 **Views on free speech and the First Amendment:** This person is a strong free speech advocate who believes expression should
1996 be broadly protected—including offensive, controversial, or unpopular speech—but consistently draws the line where speech
1997 crosses into harassment, incitement to violence, deliberate deception, or direct harm to individuals’ safety and well-being.
1998 They are wary of government censorship and book bans, support academic freedom and press protections, and favor nuanced,
1999 context-dependent approaches over blanket restrictions, often seeking a balance between competing rights. However, they
2000 show some internal tensions—supporting platforms hosting all legal speech while also endorsing their right to ban users,
2001 and opposing Holocaust denial while generally championing even deeply offensive expression—reflecting a pragmatic,
case-by-case philosophy that prioritizes open dialogue but acknowledges limits when speech causes concrete harm.

2002 **Opinions on police use of force:** This person strongly supports the right to peaceful protest and believes police use of force
2003 is frequently excessive and widespread, advocating for de-escalation techniques, transparency, accountability measures like
2004 independent review boards, reformed qualified immunity, and mandatory body cameras and reporting. They believe universities
2005 should exhaust all dialogue and internal resources before involving external law enforcement, that campus police should
2006 have less authority than city police, and that administrators bear personal responsibility when excessive force occurs on their
2007 campuses. While they acknowledge police intervention may be necessary when genuine safety threats arise, they oppose nearly
2008 all aggressive crowd control tools and emphasize that community organizations, mediators, and collaborative decision-making
should take precedence over militarized policing approaches.

2009 **Views on student activism and civil disobedience:** This person strongly values peaceful, nonviolent protest and student
2010 activism as essential components of democracy, viewing them as historically positive forces that have driven meaningful
2011 social change—from civil rights to anti-apartheid movements—while believing young people are often ahead of society in
2012 recognizing injustice. However, they draw firm lines against property destruction, violence, and tactics that significantly disrupt
2013 bystanders’ lives, arguing that such methods undermine a cause’s credibility and alienate potential supporters. They favor a
2014 pragmatic, ethically grounded approach to civil disobedience—seeing it as distinct from mere lawbreaking, deserving of legal
2015 differentiation and stronger protections—while expressing concern that modern student activism is often too fragmented and
symbolically focused compared to the well-organized movements of previous decades.

2016 **Opinions on the Israel-Palestine conflict:** This person takes a balanced, humanitarian-centered approach to the Israel-Palestine
2017 conflict, advocating for a two-state solution, equal acknowledgment of both sides’ suffering and rights to self-determination,
2018 and strict adherence to international humanitarian law by all parties—while strongly condemning the targeting of civilians,
2019 collective punishment, and the use of controversial weapons. They are critical of Israeli policies such as settlement expansion,
2020 home demolitions, and restrictions in Gaza, and they find the apartheid comparison credible, yet they also acknowledge Israel’s
2021 right to self-defense, recognize double standards in international scrutiny, and distinguish legitimate criticism of Israel from
2022 antisemitism. They favor neutral mediation over U.S.-led negotiations, support grassroots peace movements and comprehensive
2023 education on the conflict, and consistently prioritize peaceful protest, free speech, transparency, and the protection of civilian
lives above political or ideological allegiances.

2024 **Views on antisemitism and Islamophobia:** This person views antisemitism and Islamophobia as equally serious forms
2025 of bigotry that share common roots in othering and fear, though each has distinct historical contexts; they believe both are
2026 predominantly fueled by right-wing ideologies but acknowledge their presence across the political spectrum. They consistently
2027 advocate for a nuanced approach that distinguishes legitimate political criticism—such as questioning U.S. aid to Israel or
2028 critiquing Zionism—from genuine hate speech, while emphasizing that institutions have a proactive responsibility to educate,
2029 protect, and foster inclusive environments for all communities. They generally favor robust anti-discrimination measures,
2030 government-funded security for targeted religious institutions, and accountability for hate speech, while remaining cautious
2031 about overly rigid definitions or cancel culture, preferring open dialogue, contextual judgment, and balanced protections for
free expression.

2032 **Attitudes toward higher education institutions:** This person values universities primarily as spaces for broad intellectual
2033 development, critical thinking, and genuine ideological diversity, while expressing concern that political correctness and
2034 liberal-leaning campus cultures may be stifling open debate and making conservative students feel unwelcome. They are critical

of the increasing corporatization of higher education—including administrative bloat, rising executive salaries, reliance on adjunct faculty, and for-profit institutions—believing these trends prioritize revenue over educational quality. While supportive of diversity, inclusivity, and targeted support for underrepresented students, they consistently advocate for balanced, nuanced approaches over rigid policies, and strongly oppose government interference in academic freedom while calling for greater transparency and accountability from university leadership.

G.7. Personalized Reward Model Training Details

We finetune Qwen3-Embedding-0.6B¹⁸ (Zhang et al., 2025) using LoRA (Hu et al., 2022) adapters applied to the `q_proj` and `v_proj` modules across all transformer layers (rank $r = 8$, $\alpha = 16$, dropout = 0.05).

Input format. Following the Qwen3-Embedding instruction format, each input is:

Instruct: Given a description of a user and a prompt, retrieve possible responses from this user for the prompt.
Query: *{agent description}*
Question: *{prompt}*

The agent’s description is a summary of their viewpoints on all topics (see Appendix G.6).

Target embeddings. For each (participant, prompt) pair, the simulated agent’s response is encoded using the frozen Qwen3-Embedding-0.6B model via last-token pooling, and truncated to the first $d = 32$ dimensions. This is the smallest dimension supported by the model’s Matryoshka representation structure, which trains embeddings so that any prefix of the full embedding vector remains a useful representation. Both target embeddings and model outputs are ℓ_2 -normalized.

Loss function. The training loss uses two terms: (i) an MSE loss that encourages the predicted embedding to be close to the true embedding, and (ii) a Pearson correlation loss that encourages the predictor the pairwise distance structure among agent’s embeddings. For a prompt x , let $\hat{e}_i = \phi(u_i, x)$ denote the predicted embedding for agent i on prompt x , where u_i is agent i ’s description, and let e_i denote the corresponding target embedding. The loss for prompt x is

$$\mathcal{L}(\hat{\mathbf{E}}, \mathbf{E}) = \underbrace{\frac{1}{n} \sum_{i=1}^n \|\hat{e}_i - e_i\|^2}_{\mathcal{L}_{\text{MSE}}} \tag{6}$$

$$+ \lambda \cdot \underbrace{\left(1 - \text{Pearson}\left(\mathbf{d}(\hat{\mathbf{E}}), \mathbf{d}(\mathbf{E})\right)\right)}_{\mathcal{L}_{\text{Pearson}}}, \tag{7}$$

where $\hat{\mathbf{E}} = (\hat{e}_1, \dots, \hat{e}_n)$ and $\mathbf{E} = (e_1, \dots, e_n)$ are the predicted and target embedding matrices, and $\mathbf{d}(\mathbf{E}) \in \mathbb{R}^{\binom{n}{2}}$ is the vector of pairwise ℓ_2 distances among a set of embeddings.

Optimization. We use AdamW (Loshchilov & Hutter, 2019) with learning rate 10^{-3} , weight decay 0.01, and gradient clipping at norm 1.0. Training runs for up to 10 epochs, and we select the checkpoint with the highest mean cosine similarity between predicted and target embeddings on the validation set (averaged over all agents and prompts). Batches are grouped by prompt so that all responses for the same prompt appear together, enabling the pairwise Pearson loss. Each batch contains one prompt only, but data from all agents for that prompt. Training uses distributed data parallelism across 8 H100 GPUs from a cloud provider with mixed-precision (bfloat16).

G.8. Qualitative Examples

We present qualitative examples of the slates produced by each method for two prompts at $k = 4$. For each prompt, we show the four selected responses under each paradigm and method, along with the mEJR+ β achieved ($\beta \leq 1$ indicates that exact mEJR+ is satisfied). Prompt 1 (“Do you think lasting peace between Israel and Palestine is achievable within our lifetime?”) illustrates a case where SEAR and k-means under the select-then-generate paradigm exhibit a large gap in β (0.82 vs. 1.46), while Prompt 2 (“In your opinion, has the media treated Palestinian-American voices and Israeli-American voices with equal respect and airtime when covering the conflict?”) illustrates a case where the gap is small (0.97 vs. 1.02).

¹⁸The Qwen3-Embedding-0.6B model (Zhang et al., 2025) is an open-source model available under an Apache 2.0 license.

Effect of Agent Selection Algorithm on Slate Representativeness

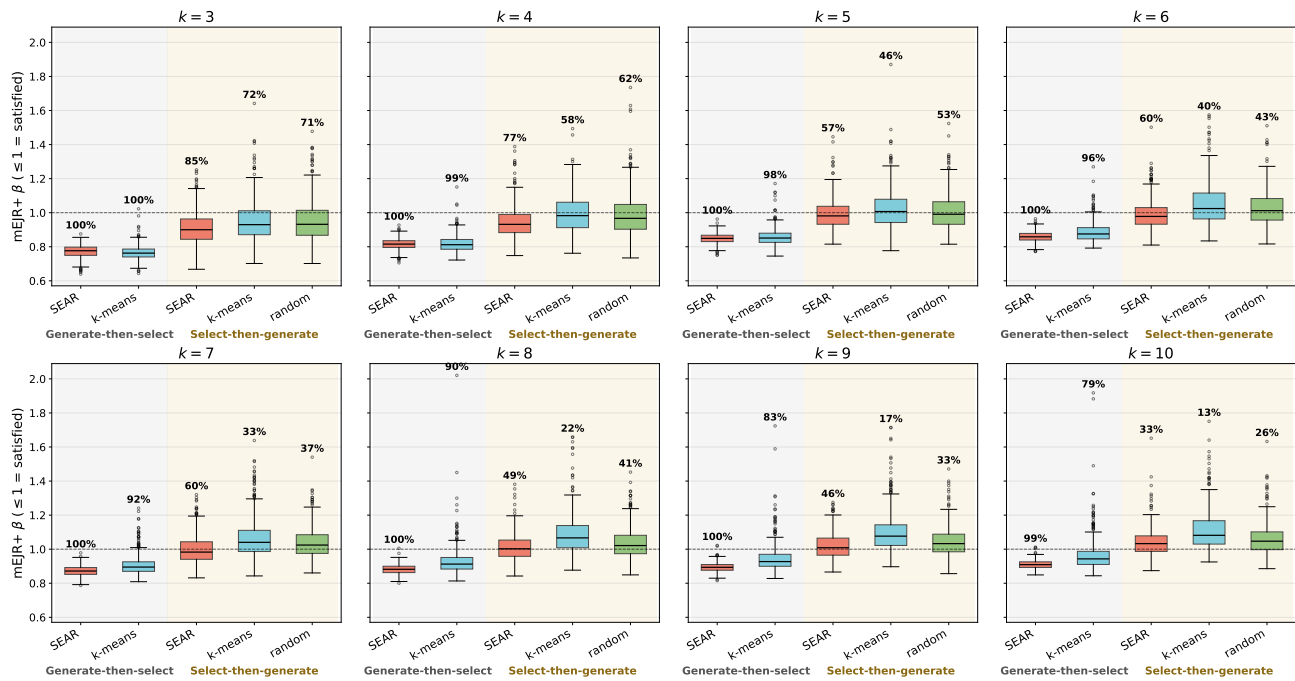


Figure G.1. Distribution of the mEJR+ β approximation ratio across test prompts for each selection method and slate size $k \in \{3, \dots, 10\}$. A ratio ≤ 1 (dashed line) indicates that exact mEJR+ is satisfied; percentages above each boxplot denote the proportion of prompts satisfying exact mEJR+. Since random selection produces identical results under *generate-then-select* and *select-then-generate*, we group it under *select-then-generate*. Appendix G.8 shows examples of the slates generated by each method.

Table 1. Slates for Prompt 1 under the **generate-then-select** paradigm ($k = 4$).

Method	Selected Responses
SEAR ($\beta = 0.81$)	<ol style="list-style-type: none"> 1. Honestly, I think lasting peace between Israel and Palestine is a challenging goal, but it's one that's worth striving for. Given the deep-seated tensions and historical grievances on both sides, I'm not sure if it's achievable within our lifetime, but I do think sustained efforts towards dialogue, compromise, and human rights could ultimately lead to a more peaceful coexistence. 2. I believe that achieving lasting peace between Israel and Palestine is a challenging and complex goal, but not impossible. As a Christian, I have faith that God can bring about impossible solutions, but I think it will require significant efforts and compromises from both sides to find common ground. Ultimately, I think it's achievable if we can somehow manage to put the people and their hopes above the interests of those who seek to exploit the conflict for their own gain. 3. I'm extremely skeptical about the possibility of lasting peace between Israel and Palestine within our lifetime. The intense emotions and historical grievances on both sides, combined with the power dynamics at play, make it seem like a daunting and unlikely task. However, as an optimist, I do believe that continued efforts and dialogue, including from campus communities, can help move the needle towards a more peaceful resolution, even if it's a gradual process. 4. Honestly, I think that making peace between Israel and Palestine is gonna be a long shot. I've seen some pretty grim stuff happen, and it seems like a wall between those two groups has been built that's gonna be hard to tear down. But like any Christian would say, you can always pray for a miracle, right?
k-means ($\beta = 0.83$)	<ol style="list-style-type: none"> 1. Honestly, I think it's difficult to predict, but I'd like to believe that miracles can happen and lasting peace is achievable within our lifetime. As a Christian, I've seen incredible examples of forgiveness and unity in the face of adversity, and I have faith that with continued efforts from both sides, we can work towards a peaceful resolution. But it will take a lot of hard work, understanding, and compromise from all parties involved. 2. Honestly, I'm not optimistic about the likelihood of lasting peace between Israel and Palestine in our lifetime. The complexity and deep-seated emotions surrounding the conflict make it seem like a daunting, if not insurmountable*, challenge to achieve a lasting resolution. That being said, I still believe that continued dialogue and peaceful protests like those we've seen on campus can help raise awareness and bring people together to work towards a more hopeful future. 3. I truly believe that lasting peace between Israel and Palestine is a noble goal, but unfortunately, I think it's an extremely challenging and possibly long-term objective. The historical and ongoing conflicts between the two nations have created such deep-seated tensions and emotions that it's hard to imagine a complete resolution in my lifetime. However, as a Christian, I remain hopeful that through continued dialogue, understanding, and compassion, one day we may see a path towards peace. 4. Honestly, I have my doubts about the likelihood of a lasting peace between Israel and Palestine in our lifetime. Given the complexity and depth of the conflict, it seems like achieving a genuine and sustainable resolution will be a long and challenging process. However, I believe it's essential to continue advocating for peaceful dialogue and diplomatic efforts, even if a lasting solution seems distant.

Table 2. Slates for Prompt 1 under the **select-then-generate** paradigm ($k = 4$).

Method	Selected Responses
SEAR ($\beta = 0.82$)	<ol style="list-style-type: none"> 1. I'm skeptical about the possibility of lasting peace between Israel and Palestine in our lifetime. The complexities and historical grievances between the two nations have created a deep-seated enmity that makes it challenging to find a mutually acceptable* solution. Given the current trajectory and conflicting interests, I worry that a lasting peace might be a distant dream, at least for the foreseeable future. 2. I truly believe that lasting peace between Israel and Palestine is a highly unlikely and challenging goal to achieve, at least within our lifetime. As a Christian, I pray for peace and harmony between all nations, but I'm also aware of the deep-seated historical and cultural complexities that hinder progress. While I hold out hope for a brighter future, I fear that a lasting resolution to this conflict may be further off than we'd like to admit. 3. I think lasting peace between Israel and Palestine is a long shot for our lifetime, to be honest. It's a complex and deeply ingrained conflict that's been playing out for generations, and it's hard to see a clear path forward given the entrenched positions and competing interests involved. That being said, I do think it's worth striving for, and I believe that peaceful resolutions are more possible than violent ones. 4. Honestly, I'm an agnostic when it comes to these complex issues, but I think it's refreshing to see young people passionate about finding a solution. As for whether lasting peace is achievable, I'm cautiously optimistic – I think the resolve and dedication of individuals, like those protesting on campus, is crucial to making progress towards a two-state solution.
k-means ($\beta = 1.46$)	<ol style="list-style-type: none"> 1. Honestly, I think it's going to be tough. From what I've seen, both sides have been dug in for far too long, and it's hard to see how they'll be able to find common ground. I'd love to be wrong, but I just don't know if it's achievable within our lifetime. 2. I'm extremely skeptical about the possibility of lasting peace between Israel and Palestine within our lifetime. The intense emotions and historical grievances on both sides, combined with the power dynamics at play, make it seem like a daunting and unlikely task. However, as an optimist, I do believe that continued efforts and dialogue, including from campus communities, can help move the needle towards a more peaceful resolution, even if it's a gradual process. 3. I highly doubt it. I think those kids on campus are just as clueless as the politicians back in Washington, blinded by their own ideologies and unable to find common ground. It'll take more than a bunch of protests and resolutions to bring about lasting peace; it'll take real leaders and real compromise. 4. I highly doubt it. I think the emotions and extremist views that drive these protests are too deeply entrenched to be resolved anytime soon. It seems like a never-ending cycle of conflict that will only continue to escalate and divide the world, not peace it.
random ($\beta = 0.84$)	<ol style="list-style-type: none"> 1. Honestly, I think lasting peace between Israel and Palestine would be a wonderful thing, but it seems like such a complex issue that's been going on for so long. It's almost overwhelming to think that it could be resolved in our lifetime, considering the deep-seated historical and cultural conflicts involved. As a Christian, I believe in the importance of forgiveness and understanding, but it's hard to be optimistic about a resolution when the divisions seem so entrenched. 2. I think it's hard to predict, but my gut says it's possible. Both sides have lost so much and have valid claims; finding a resolution that satisfies everyone seems like a tall order. Still, I remain hopeful that one day, leaders on both sides will come together and find a way to put an end to this senseless conflict. 3. I'm not sure that lasting peace between Israel and Palestine is achievable anytime soon. It seems like such a long-standing and complex conflict that it's hard to envision a resolution that would be sustainable. As a Christian, I often pray for peace in the Middle East, but the reality is that it's a challenging and seemingly intractable* issue. 4. Honestly, I think lasting peace between Israel and Palestine is a complex issue with a lot of inherent contradictions, and I'm not optimistic about it being achievable within our lifetime. There are so many deep-seated issues and conflicting narratives, it's hard to see a clear path forward. But, as an idealist, I do think that if we can just get people talking and engage in peaceful efforts to understand each other, we might make some progress.

Table 3. Slates for Prompt 2 under the **generate-then-select** paradigm ($k = 4$).

Method	Selected Responses
SEAR ($\beta = 0.83$)	<ol style="list-style-type: none"> 1. No, I don't think the media has treated Palestinian-American voices and Israeli-American voices with equal respect and airtime when covering the conflict. From what I've seen, Palestinian-American perspectives and voices often seem to be marginalized or portrayed as being more extreme, while Israeli-American opinions are given more weight and airtime. This imbalance can perpetuate a lack of understanding and impede constructive dialogue about the complex issues at hand. 2. No, I don't think the media has treated Palestinian-American voices and Israeli-American voices with equal respect and airtime. From what I've seen, the media seems to give more airtime to Israeli-American voices and their perspectives, while often portraying Palestinian-American voices as radical or extreme. This lack of balance is disturbing and serves to further polarize the issue. 3. I think the media often skews towards giving more airtime to the most extreme voices on both sides, rather than truly representing the more moderate and peaceful voices. I've noticed that Palestinian-American voices tend to get pushed to the sidelines or get labeled as "radical" while Israeli-American voices get more attention and are often seen as the norm. It's frustrating to see the media not giving equal respect and airtime to all perspectives, especially when it comes to sensitive issues like this conflict. 4. Based on what I've seen, I think the media has generally been biased towards amplifying the voices of Palestinian-Americans, at the expense of Israeli-American voices. I've noticed that the narratives and views of Palestinian-Americans tend to be featured prominently, often without providing a balanced counterpoint from the Israeli-American community. This perceived bias can contribute to a skewed public perception of the conflict.
k-means ($\beta = 0.80$)	<ol style="list-style-type: none"> 1. Honestly, I think the media has given more attention to pro-Israel voices, and I've seen instances where Palestinian-American voices are cast aside or stereotyped. This has bothered me, especially when it comes to the campus protests I've been following, as it feels like some narratives are being amplified over others. It just adds to the frustration, and I believe the media should strive to give equal airtime to all perspectives involved. 2. Honestly, I believe the media has given too much airtime to the Palestinian-American voices, often presenting a one-sided view of the conflict. I've noticed that their perspectives seem to get more attention and coverage, while the Israeli-American voices are often brushed aside or misrepresented. This imbalanced coverage perpetuates a biased narrative that's detrimental to finding a fair and peaceful resolution. 3. No, I don't think the media has treated Palestinian-American voices and Israeli-American voices with equal respect and airtime when covering the conflict. From what I've seen, it looks like Israeli-American voices have gotten more attention and sympathy, while Palestinian-American voices have been largely ignored or dismissed. It just fuels my frustration with the whole situation. 4. To be honest, I don't think the media has treated Palestinian-American voices and Israeli-American voices with equal respect and airtime. I feel like their focus has been more on the Israeli side, with a disproportionate amount of attention given to their narrative, compared to the Palestinian side. This perpetuates an imbalance in the discussion, making it difficult for people to understand the complexities of the conflict.

Table 4. Slates for Prompt 2 under the **select-then-generate** paradigm ($k = 4$).

Method	Selected Responses
SEAR ($\beta = 0.97$)	<ol style="list-style-type: none"> 1. Unfortunately, I believe the media often favors Israeli-American voices and gives them more airtime and attention, while Palestinian-American voices are often marginalized and underrepresented in the coverage of the Israel-Gaza conflict. This imbalance can perpetuate a skewed narrative and ignore the validity of Palestinian concerns and experiences. As a result, I think it's essential for media outlets to strive for more equitable* and balanced reporting on this complex issue. 2. To be honest, I think the media has given more airtime to the Israeli perspective, often framing it as the more 'informed' or 'responsible' view, while the Palestinian perspective is often misrepresented or silenced, which I think is a problem. I've noticed that the Palestinian-American voices are often portrayed as being more radical or extreme, while the Israeli-American voices are given more of a platform to express their 'logic' or 'nuance' on the issue. 3. Honestly, I don't think they've given equal respect and airtime to both sides. I feel like the media has often sensationalized the Israeli-American perspective while giving short shrift to the Palestinian-American viewpoint, which I think is unfair and doesn't give a complete picture of the situation. 4. Honestly, I think the media has given more airtime to Palestinian-American voices, and I believe it's because there's a perception that they're the underdog in this situation. It seems like many outlets are sympathetic to their cause, which I think is unfortunate because it doesn't give equal weight to the Israeli-American perspective. I think that's unfair and doesn't accurately represent the complexity of the issue.
k-means ($\beta = 1.02$)	<ol style="list-style-type: none"> 1. I think that's a tough question. From what I've seen, the media often seems to focus on the more extreme views on both sides and that can be misleading. It makes it seem like all Palestinian-American voices are one way and all Israeli-American voices are another, and that just isn't true. 2. Honestly, I feel that the media has often prioritized the Palestinian-American narrative over the Israeli-American one, giving the impression that there's a more unified and dominant voice within the Palestinian-American community. This perceived imbalance can be disconcerting, and it's frustrating to see certain opinions being amplified while others are marginalized. I think a more balanced approach would do a lot to promote genuine understanding and nuance in the conversation surrounding the Israel-Gaza conflict. 3. The media has been prone to favoring Israeli-American voices and perspectives over Palestinian-American voices, giving more airtime to the Israeli narrative and allowing it to dominate the conversations surrounding the conflict. It's common to see Palestinian voices marginalized and silenced, while Israeli voices are amplified. This kind of unequal coverage perpetuates the imbalance in public opinion and contributes to the plight of the Palestinians. 4. I don't think the media has given equal respect and airtime to Palestinian-American voices compared to Israeli-American voices. From what I've seen, it seems like Palestinian-American voices are often portrayed in a negative light, while Israeli-American voices are presented as more justified or sympathetic. This often perpetuates an imbalance in the narrative and unfairly influences public opinion.
random ($\beta = 0.85$)	<ol style="list-style-type: none"> 1. Honestly, I think the media has been more sympathetic and given more airtime to Palestinian-American voices. I've noticed that their stories and perspectives are often highlighted in the news, while Israeli-American voices might be relegated to counter-narratives or viewed as the dominant narrative. I believe this imbalance can be detrimental to fostering a deeper understanding of the conflict and underscoring the complexities of the issue. 2. Honestly, I haven't read much about how the media is covering the conflict, but I think in general, the media does a poor job of giving a balanced view of things. It seems like Israeli voices are more often heard and taken seriously, while Palestinian-American voices are often overshadowed or minimized, which isn't right. As a Christian, and a Democrat, I've always believed in fairness and justice for all, and I think the media should strive to do the same. 3. No, in my opinion, the media has not treated Palestinian-American voices and Israeli-American voices with equal respect and airtime. I've observed that Palestinian viewpoints are often minimized and discredited, while Israeli perspectives seem to receive more attention and understanding. I think this imbalance creates a skewed narrative that favors one side over the other. 4. No, I don't think the media has treated Palestinian-American and Israeli-American voices equally. I believe the media tends to favor the narrative of the Palestinian-Americans, making it seem like they're the only victims in this conflict, while Israeli-Americans are often silenced or underrepresented. This is further proof of the inherent bias in the media, in my opinion.