Social Choice Foundations for Simulation-Augmented Generation

Sonja Kraiczy [1 2 *] **Smitha Milli** [3 *] **Ratip Emin Berker** [4] **Avinandan Bose** [3 5] **Brandon Amos** [3]
Jamelle Watson-Daniels [3] **Maximilian Nickel** [3] **Edith Elkind** [6] **Ariel D. Procaccia** [3 7]

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 investigate how the guarantees achieved by our method are affected by the fidelity between the simulation and the human.

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: 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, Milli et al. (2026) 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) (Milli et al., 2026): at inference time, given a user prompt, the model queries generative simulations of individuals in a target population—*simulants*—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. Standard approaches to satisfying these constraints may interfere with representation. For example, naively summarizing simulant 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**.

[*]Equal contribution [1]University of Oxford [2]Principles of Intelligence [3]Meta Superintelligence Labs [4]Carnegie Mellon University [5]University of Washington [6]Northwestern University [7]Harvard University. Correspondence to: Sonja Kraiczy <sonja.kraiczy@cs.ox.ac.uk>, Smitha Milli <smilli@meta.com>.

AI4GOOD Workshop @ ICML 2026, Seoul, South Korea. Copyright 2026 by the author(s).

1.1. Overview of Approach

Our research question is this: given a population of m humans (e.g., millions), can we learn only n *simulants* (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 simulants 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 simulants 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, simulants generate open-ended responses to a large and essentially unconstrained set of prompts. Preferences are therefore most naturally represented at the individual level via reward functions: each human i has a reward function $r_i(p, o)$ that yields their utility to a response o for prompt p.

The representation guarantee we then seek is: given a prompt p, we aim to select k simulants to query so that their responses *proportionally represent* the viewpoints of the overall population of humans. Informally, when a large group of humans shares similar utilities for responses to this prompt, then a proportionate fraction of the k simulated responses should deliver high utility to this group. Moreover, the utility level that is guaranteed to the group depends on how cohesive their preferences are: groups with more aligned preferences can be guaranteed higher utility.

Efficiently guaranteeing representation. We prove that to represent a population of m humans, we need only simulate a random sample of $n \ll m$ individuals where n grows logarithmically in m. We further prove that at inference time, querying just $k \ll n$ prompt-dependent simulants suffices to achieve approximate representation guarantees. To accomplish this, we adapt proportional clustering algorithms to our setting and use them to dynamically select k simulants given a prompt. The degree of representation achieved by their responses depends on the fidelity gap between all simulants and their humans.

1.2. Our Contributions

Below we state our technical contributions.

1. **Formal representation guarantees.** We provide the first formalization of representative simulation in SAGE through the careful adaptation of axioms from the computational social choice literature, chiefly *proportionally representative fairness* (Aziz et al., 2024).

2. **Efficient proportional clustering.** We develop a faster implementation of the spatial expanding approval rule (SEAR)[1], 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, which makes it practical for inference-time deployment.

3. **Efficient querying of simulations.** Under the natural assumption of a factorized reward structure across individuals (akin to collaborative filtering), we prove that an adaptation of proportional clustering yields an efficient inference-time procedure that only requires querying $k \ll n$ simulants for a given prompt. The responses from these k simulants are approximately representative, with the level of approximation depending on the fidelity gap between simulants and their corresponding humans.

4. **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 prove that it suffices to run our method on a population of $n \ll m$ simulants that simulates a random (possibly stratified) sample of the true population: specifically, n need grow only logarithmically in m, and when the reward model dimension d is small, n can depend only on d and be independent of m.

5. **Empirical validation.** We evaluate our approach against baselines through semi-synthetic experiments using data from a deliberative discussion platform, in which we synthetically control the fidelity gap between simulants and humans to assess representativeness as a function of fidelity. Across all fidelity gaps, our theoretically-motivated inference-time algorithm achieves the best representation, with the most substantial gains in the high-fidelity regime.

2. Related Work

Social choice. Starting with the work of Aziz et al. (2017), an influential line of research in social choice formalizes the notion of proportional representation in committee elections and participatory budgeting (Aziz et al., 2018; Peters et al., 2021). Generally speaking, the notions introduced therein ensure that every subset of simulants that is sufficiently large, and whose preferences are sufficiently cohesive, would be represented in the outcome. Importantly, a similar approach

[1] More precisely, we consider a variant of this algorithm, discussed in Section 3.

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 simulants' 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 mPJR+, 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). Their goal, too, is to generate a slate of viewpoints that is formally representative of the views of a given set of simulants. 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 inference-time scalability criterion.

In general, most works on social choice for AI alignment focus on aggregation of preferences or rewards prior to deployment (Ge et al., 2024; Conitzer et al., 2024; Gölz et al., 2025; Hollender & Kraiczy, 2026; Maura-Rivero et al., 2025), rather than at inference time as our approach does. A recent exception is concurrent work by Freedman (2026), who propose generating diverse candidate responses at inference time and then applying a social choice voting procedure with heterogeneous reward functions to select the winner.

Pluralism in LLMs. SAGE (Milli et al., 2026) 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 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 simulants while enabling adaptive, prompt-specific selection from a larger population of n simulated individuals, with formal guarantees that the selected simulants proportionally represent the broader population.

3. Proportional Clustering

In our work, we adapt concepts from the proportional clustering literature to formalize the problem of representative simulation. We first begin by describing these concepts in their original form in an abstract geometric model of preferences. Following spatial models of preferences (Coombs, 1964; DeSarbo & Hoffman, 1987; Poole & Rosenthal, 1985), suppose that, for a fixed prompt, simulants are represented by ideal points $v_1, \ldots, v_n \in \mathbb{R}^d$, with distances encoding preferences. This gives an idealized way to think about proportional polling in a large, continuous space of possible viewpoints: if we could directly select viewpoints, we would want a slate $X \subseteq \mathbb{R}^d$ of size k that proportionally represents the population.

Representation Guarantees. Let $\mathcal{N}$ be the set of simulants, with $|\mathcal{N}| = n$, and let Δ denote Euclidean distance. For $\ell \in [k]$, we say that a group $S \subseteq \mathcal{N}$ is *ℓ-large* if $|S| \geq \ell n/k$; such a group has a proportional claim to ℓ of the k slate positions. Size alone, however, is not enough to determine representation: a large but widely dispersed group should not have the same claim to nearby representatives as a large, tightly clustered group. We therefore also measure cohesiveness. For $S \subseteq \mathcal{N}$, define $D(S) := \max_{i,j \in S} \Delta(v_i, v_j)$, and say that S is *t-cohesive* if $D(S) \leq t$. Smaller values of t correspond to more cohesive groups, which should be represented by slate points closer to their members.

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

Definition 1 (Proportionally Representative Fairness)**.** We say that a slate X satisfies *proportionally representative fairness* (PRF) if, for every $\ell \in [k]$, $t \in \mathbb{R}$ and every ℓ-large and t-cohesive group of simulants $S \subseteq \mathcal{N}$, it holds that

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

Efficient and Proportional Clustering. To select k points satisfying our representation guarantees, we use a *proportional* clustering algorithm— a variant of *spatial expanding approval rule* (SEAR).[2] Unlike (balanced) k-means, SEAR

[2]The 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.

provably returns k centroids satisfying PRF when the candidate centroids are exactly the voter points.

SEAR. The input to SEAR is a multiset of q candidate centroids $C = (c_1, \ldots, c_q)$, a multiset of n voters $V = (v_1, \ldots, v_n)$, and a target slate size k. We treat candidates and voters as indexed objects, so distinct candidates or voters may occupy the same point in space. Each voter initially has budget $b_i = 1$, and the slate is initialized as the empty multiset X. The algorithm grows Euclidean balls around candidate centroids at uniform speed. As soon as the ball around some candidate centroid c_j contains voters $S \subseteq V$ whose total remaining budget is at least n/k, the budget of voters in S is reduced by n/k in total, and c_j is added to X. The algorithm terminates and returns X once $|X| = k$.

The PRF guarantee applies in the special case where the candidate centroids are exactly the voter points, i.e., $q = n$ and $c_i = v_i$ for each $i \in [n]$.

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 related precursor of this algorithm 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 1. *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 $b_v = n/k - \lfloor n/k \rfloor$.

4. Selection and Generation Commute

The previous section gave a guarantee for selecting k representative points among a larger set of points in Euclidean space. We now connect this guarantee to inference-time generation, where the slate must consist of text responses rather than points. Throughout this section, we assume access to two objects: for each human, their reward function over responses, and the simulant that generates responses on their behalf. These objects may be learned in practice; here we treat them as primitives.

Inference-Time Social Choice Fix a prompt p and a response space $\mathcal{V}^*$. For each human $i \in \mathcal{N}$, let $r_i(p, \cdot) : \mathcal{V}^* \to \mathbb{R}$ denote human i's reward function over possible responses to p. Let A_i denote the corresponding simulant, so that $A_i(p) \in \mathcal{V}^*$ is the response generated by simulant i on prompt p.

Ideally, A_i is perfectly aligned with human i: it generates a response, or viewpoint, $o_i^* \in \arg\max_{o \in \mathcal{V}^*} r_i(p, o)$ that maximizes i's reward function. More generally, we allow $A_i(p)$ to be an approximate maximizer of $r_i(p, \cdot)$, so that the simulant's response approximately reflects the viewpoint human i would endorse.

For a group $S \subseteq \mathcal{N}$, define its inference-time cohesiveness by $\tau(S) := \min_{i,j \in S} r_i(p, A_j(p))$. We say that S is τ-cohesive on prompt p if $\tau(S) \geq \tau$.

Definition 2 (Inference-Time PRF)**.** A slate $Y \subseteq \mathcal{V}^*$ of k responses satisfies *inference-time proportionally representative fairness* (IT-PRF) on prompt p if, for every ℓ-large and τ-cohesive group $S \subseteq \mathcal{N}$, it holds that

$$\left| \left\{ o \in Y \;\middle|\; \max_{i \in S} r_i(p, o) \geq \tau \right\} \right| \geq \ell.$$

The definition above is stated in terms of rewards over responses, while the selection procedure from the previous section operates on points in a Euclidean space. To reconcile these two frameworks, we introduce the following assumption regarding the structure of the reward function.

Assumption 1 (Factorized reward)**.** We assume that the reward function admits a factorized structure that is shared across individuals:

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

where $\psi : \mathcal{V}^* \to \mathbb{S}^{d-1}$ and $\phi_i : \mathcal{V}^* \to \mathbb{S}^{d-1}$ for $i \in \mathcal{N}$, so $\psi(o)$ embeds responses and $\phi_i(p)$ is simulant i's prompt-specific embedding.

This suggests a conceptually complete baseline that applies the proportional clustering procedure from the previous section.

Generate-then-Select. Given a prompt p, generate each simulant's response $o_i = A_i(p)$. Let $C = (\psi(o_i))_{i \in \mathcal{N}}$ be the candidate response embeddings and $V = (\phi_i(p))_{i \in \mathcal{N}}$ the voter embeddings. Run SEAR with candidates C and voters V to select k representative responses indexed by $R \subseteq \mathcal{N}$, and return

$$Y := \{o_i : i \in R\}.$$

Generate-then-Select realizes the entire simulated population and reduces inference-time representation to selection over an explicit response set: each simulant contributes

one candidate response, and the procedure then selects k of these responses. After generation, this is simply a representative selection problem over a fixed response set with simulant-specific rewards.

Theorem 2. *Generate-then-Select returns a slate Y that satisfies IT-PRF.*

Select-then-Generate

Input: prompt p, simulants $\mathcal{N}$, reward functions $\{r_i\}_{i\in\mathcal{N}}$, simulation models $\{A_i\}_{i\in\mathcal{N}}$, slate size k.
Output: slate Y of k generated viewpoints.

1. **Compute embeddings.** Compute $\{\phi_i(p)\}_{i\in\mathcal{N}}$ for prompt p, where $r_i(p,o) = \phi_i(p)^\top \psi(o)$.
2. **Select k representatives.** Run SEAR on $\{\phi_i(p)\}_{i\in\mathcal{N}}$ to obtain the embeddings $\{\phi_i(p)\}_{i\in R}$ of a set $R \subseteq \mathcal{N}$ of k simulants.
3. **Query selected simulants.** Generate $A_i(p)$ *only* for simulants $i \in R$.
4. **Return slate.** Return $Y \leftarrow \{A_i(p)\}_{i\in R}$.

The guarantee comes at a steep computational cost: Generate-then-Select requires one generation per simulant. When the simulators A_i are large autoregressive models, especially ones using inference-time reasoning, this full-population generation step becomes prohibitive in terms of both computational resources and latency. The goal is therefore to retain the same representational benefits with greater scalability.

This suggests a useful commutation trick: reverse the order of selection and generation. The resulting procedure is *Select-then-Generate*: Instead of generating one response per simulant and then selecting representatives, we first select representative simulants in the prompt-conditioned embedding space and then generate responses only for the k selected simulants. As the next theorem shows, under faithful simulation, the costly simulations of unselected simulants can be skipped without losing proportional representation: Any large, cohesive group that would be represented by generating all simulants is already represented after selecting in the prompt-conditioned embedding space and generating only the selected responses.

We say that A_i is *faithful* on prompt p if it outputs a response that realizes maximal reward for human i. Under the factorized reward assumption (Assumption 1), this implies that

$$\psi(A_i(p)) = \phi_i(p).$$

If the simulations are faithful, Select-then-Generate coincides exactly with Generate-then-Select, resulting in the following theorem.

Theorem 3 (Exact Selection–Generation Commutation)**.** *Fix a prompt p, and suppose every simulator A_i, $i \in \mathcal{N}$, is faithful on p. The slate Y returned by Select-then-Generate satisfies IT-PRF.*

5. Approximate Simulation and Population Guarantees

In general, simulations need not be perfectly faithful to the corresponding human's viewpoint. We capture this approximation by a fidelity gap: simulation A_i has *fidelity gap at most* α on prompt p if $r_i(p, A_i(p)) \geq \cos(\alpha)$. Since a nonzero fidelity gap can prevent exact IT-PRF from holding, we define the following approximate version. We say that a slate $Y \subseteq \mathcal{V}^*$ of k responses satisfies $(\varepsilon, \gamma_\beta)$*-approximate IT-PRF* on prompt p if, for every group $S \subseteq \mathcal{N}$ with $|S| \geq (\ell/k + \varepsilon)n$ such that $\tau(S) \geq \cos(\beta)$, it holds that

$$\left|\left\{ o \in Y \;\middle|\; \max_{i\in S} r_i(p,o) \geq \cos(\gamma_\beta) \right\}\right| \geq \ell.$$

Here, γ_β denotes the weakened representation threshold associated with groups that are cohesive at level $\cos(\beta)$; in general, γ_β may depend on β as well as on an upper bound on the fidelity gap of the simulations. When $\varepsilon = 0$, we simply write γ_β-approximate IT-PRF.

Theorem 4 (Approximate Selection–Generation Commutation)**.** *Fix a prompt p, and suppose every simulator has fidelity gap at most α on p. Let Y be the slate returned by Select-then-Generate. Then Y satisfies γ_β-approximate IT-PRF for $\gamma_\beta = \alpha + \beta + \min\{\alpha, \beta\}$.*

Thus, the exact commutation guarantee for Theorem 3 extends to approximately faithful simulations: when simulators have fidelity gap at most α, the returned slate still satisfies approximate IT-PRF while querying only k simulations.

This leaves a second scaling question: how large must the simulant pool $\mathcal{N}$ be? Although Select-then-Generate queries only k simulants at inference time, it still assumes access to embeddings $\{\phi_i(p)\}_{i\in\mathcal{N}}$ for all simulants in the pool. If $\mathcal{N}$ contained one simulant for every member of the target population $\mathcal{M}$, then representing a population such as the United States would require maintaining hundreds of millions of simulants, which is impractical both computationally and operationally.

Fortunately, it is not necessary to instantiate a simulant for every person in the population: for representation guaran-

tees, it suffices to simulate only a much smaller, random sample of them.

We first consider groups with simple geometric structure. Call $S \subseteq \mathcal{M}$ a *cluster* if $S = \{i \in \mathcal{M} : \Delta(v_i, x) \leq \tau\}$ for some $x \in \mathbb{S}^{d-1}$ and $\tau > 0$. For such groups, PRF transfers from the sample to the population with no additional loss in the cohesiveness parameter t.[3]

Proposition 1. *Let $\mathcal{N} \subseteq \mathcal{M}$ be a uniformly random n-element subset, where $n = \Omega(\varepsilon^{-2}(d\log(1/\varepsilon) + \log(1/\delta)))$. If X^* is a PRF slate with respect to $\{v_i\}_{i\in\mathcal{N}}$, then w.p. at least $1-\delta$, for every t-cohesive cluster S of size at least $(\ell/k + \varepsilon)m$, $\ell \in [k]$, $|\{x \in X^* \mid \min_{i\in S} \Delta(x, v_i) \leq t\}| \geq \ell$.*

For arbitrary groups, PRF still transfers from the sampled humans to the full population, but the approximation introduces a geometric loss.

Lemma 1. *Let $\mathcal{N} \subseteq \mathcal{M}$ be a uniformly random n-element subset, where $n = \lceil \varepsilon^{-2}\log(km/\delta)\rceil$. If X^* satisfies PRF with respect to $\{v_i\}_{i\in\mathcal{N}}$, then w.p. at least $1-\delta$, X^* satisfies $(\varepsilon, 2)$-PRF[4] with respect to $\{v_i\}_{i\in M}$.*

Taking $m = 8 \cdot 10^9$, $k = 10$, $\delta = 0.05$, and $\varepsilon = 0.05$ gives a required sample size of roughly 1.1×10^4. Thus the theorem removes the linear dependence on the population size, though the required number of simulants may still be much larger than the slate size k. In Appendix D, we show that the arbitrary-groups guarantee can also be stated with a dimension-dependent sample size: when the embedding dimension d is small, the required sample size can be independent of both m and k.

Together, sampling and Select-then-Generate let us represent a large population using a manageable simulant pool and only k inference-time simulation queries, while preserving approximate representation for the fixed prompt p.

Theorem 5. *Let $\mathcal{N} \subseteq \mathcal{M}$ be a uniformly random n-element subset, where $n = \lceil \varepsilon^{-2}\log(km/\delta)\rceil$. Fix a prompt p, and suppose every simulator A_i, $i \in \mathcal{N}$, has fidelity gap at most α on p. Let Y be the slate returned by Select-then-Generate. Then, w.p. at least $1-\delta$, Y satisfies $(\varepsilon, \gamma_\beta)$-approximate IT-PRF with respect to $\mathcal{M}$, where $\gamma_\beta = \alpha + 2\beta + 2\min\{\alpha, \beta\}$.*

Thus, for a fixed prompt p, a sample of size logarithmic in m suffices for a high-probability representation guarantee over the full population, while Select-then-Generate uses only k simulation queries at inference time. This shifts the computational burden from maintaining simulations for everyone to constructing a much smaller pool of high-fidelity simulants.

6. Experiments

Overview. Our goal is to enable a critical component of simulation-augmented generation: routing queries to simulants in a way that is both *scalable* and *representative*. Thus far, we have developed the necessary theoretical components. In this section, we now empirically measure the level of representativeness attained by our proposed *Select-then-Generate* procedure as a function of the simulation fidelity γ between the simulants and their corresponding human.

Simulants. In simulation-augmented generation, we seek to query simulants for *open-ended* viewpoints. We therefore base our simulants 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 $n = 310$ representative participants[5] shared their viewpoints on the campus protests of 2024.[6] We construct one simulant per participant. Each simulant 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 profile[7], along with their written viewpoints and binary agree/disagree votes on other participants' opinions as few-shot context (see Appendix F.5 for the full prompt).

Fidelity γ. While these simulants are constructed by following the common practice of in-context prompting with demographic and attitudinal profiles (Hewitt et al., 2024; Cui et al., 2025; Argyle et al., 2023), we do not assume they are perfectly faithful representations of their human counterparts (Wang et al., 2025; Bisbee et al., 2024). Instead, we explicitly evaluate how the fidelity γ affects the level of representation achieved by our method. We control γ semi-synthetically via constructing a human reward function $r_i^{(\gamma)}(p, o) = \phi_i^{(\gamma)}(p)^\top \psi(o)$ such that the reward that human i receives from their simulant's response $A_i(p)$ is $r_i^{(\gamma)}(p, A_i(p)) = \gamma$.

[3] More generally, the same conclusion holds for any bounded-VC range family by uniform convergence.

[4] Here, $(\varepsilon, 2)$-PRF means that every group $S \subseteq \mathcal{M}$ with $|S| \geq (\ell/k + \varepsilon)m$ and $D(S) \leq t$ receives at least ℓ slate points within distance $2t$ of some member. The proof shows the stronger intermediate claim that some $i \in S \cap \mathcal{N}$ has ℓ slate points within distance $2t$.

[5] Participants 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 D also bound how representation guarantees extend beyond the n participants to the U.S. population more broadly.

[6] The data are available at `https://github.com/akonya/polarized-issues-data` under a CC-By-4.0 license.

[7] Gender, age, political affiliation, religious affiliation, education, urbanicity, and household income.

For the response embedding ψ, we use the embedding model that Blair et al. (2026) trained on this Remesh session. The model is fine-tuned so that distances in ψ-space predict participants' preferences rather than merely reflecting surface-level similarity: a participant i is more likely to agree with responses that are closer to their own embedded response. To simulate different levels of fidelity, we randomly perturb each simulant's response embedding $\psi(A_i(p))$ to obtain a prompt-preference embedding $\phi_i^{(\gamma)}(p)$ such that human i receives a reward $r_i^{(\gamma)}(p, A_i(p))$ of γ from their simulant's response (for details, see Appendix F.6). We then evaluate the representation level attained as a function of γ.

Verifying representation. While our theoretical results focus on the PRF axiom, unfortunately, in Theorem E.1, we prove that it is NP-hard to verify PRF. This leads us to adopt an alternative axiom for evaluation, mPJR+ (defined formally in Appendix F.1). While mPJR+ is closely related to PRF, and indeed implies discrete PRF[8], it also has the benefit of a known fixed-parameter tractable verification algorithm (Algorithm 1, He et al. (2026)). For evaluation, we generate 1,000 new prompts asking for an open-ended viewpoint related to the original questions in the session (see Appendix F.3 for details), and measure the percentage of prompts for which each method satisfies mPJR+.

Methods. We compare the representation levels attained by the following four methods.

1. *Generate-then-Select, (oracle).* This oracle method, which queries all n simulants, always satisfies mPJR+. It runs SEAR using the generated response embeddings $\{\psi(A_i(p))\}$ as candidates and prompt-preference embeddings $\{\phi_i(p)\}$ as voters.
2. *Select-then-Generate, SEAR (ours).* Our proposed method runs SEAR with the prompt-preference embeddings $\{\phi_i(p)\}$ as both the voters and candidates to select which k simulants to query.
3. *Select-then-Generate, k-means.* This baseline method runs k-means on the prompt-preference embeddings $\{\phi_i(p)\}$ and selects the k simulants closest to the cluster centers.
4. *Select-then-Generate, random.* This baseline selects k simulants uniformly at random.

Results. We evaluate all methods across the 1000 prompts for $k \in \{3, \ldots, 10\}$ and $\gamma \in \{0.50, 0.55, \ldots, 1.0\}$. As shown in Figure 1, our method (Select-then-Generate, SEAR) performs better or on par with both baselines (Select-then-Generate, k-means and random) across all γ and k. As the fidelity γ decreases, its advantage degrades: SEAR and k-means first converge toward each other, and with sufficiently poor fidelity, both become comparable to random selection.

For larger k, mPJR+ is harder to satisfy, as reflected by the lower rate of satisfaction by random selection on higher k. Intuitively, larger k induces finer-grained representation requirements—smaller groups of size n/k become entitled to representation—making it less likely that random sampling covers all such groups. Consistent with this, in the high-fidelity regime, SEAR's advantage over k-means is larger for higher k. However, the impact of fidelity also becomes more pronounced at higher k, and SEAR's advantage diminishes more quickly as γ decreases, reflecting the greater sensitivity of fine-grained representation to fidelity.

Overall, our results indicate that SEAR is the preferable selection algorithm for the *Select-then-Generate* paradigm, as it always performs better or on par with k-means and random selection. However, the results also suggest that the performance benefits from SEAR will be most notable when the simulants faithfully reflect their humans (higher γ) and when finer-grained representation is desired (larger k).

7. 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. To systematically study the impact of simulation fidelity γ, we utilized a semi-synthetic experimental set-up where we controlled γ explicitly, but this also means that it remains to be seen what levels of γ are expected in the wild, although we generally expect fidelity to improve as simulation methods improve. 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, in practice, different simulants may be more reliable in different domains (e.g., politics versus interpersonal advice). How should routing account for heterogeneous fidelity? Finally,

[8] For the distinction between discrete and unconstrained PRF, we refer the reader to Definition 4 and Definition 5 of Aziz et al. (2024).

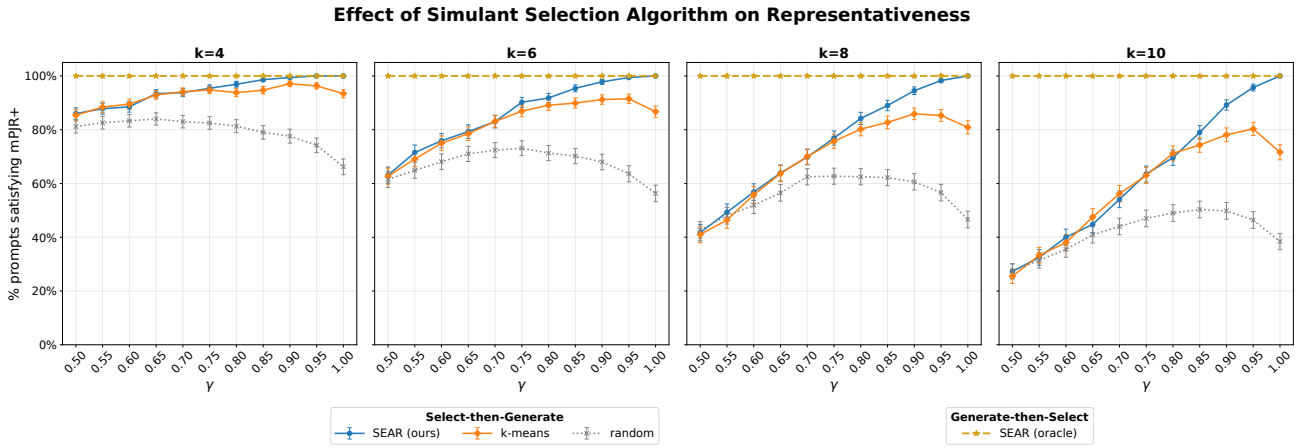

Figure 1. **mPJR+ satisfaction rate as a function of fidelity γ and the number of selected simulants k.** Results for all $k \in \{3, \ldots, 10\}$ are shown in Figure F.1.

while this work focused on the problem of routing queries to simulants, there remain several other research directions required for achieving the full vision of SAGE outlined by Milli et al. (2026), including developing higher-fidelity simulations, doing pluralistic 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 simulants 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.

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., 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.

Aziz, H., Elkind, E., Huang, S., Lackner, M., Sánchez-Fernández, L., and Skowron, P. On the complexity of extended and proportional justified representation. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*, pp. 902–909, 2018.

Aziz, H., Lee, B. E., Chu, S. M., and Vollen, J. Proportionally representative clustering. In *Proceedings of the 20th Conference on Web and Internet Economics*, 2024.

Bisbee, J., Clinton, J. D., Dorff, C., Kenkel, B., and Larson, J. M. Synthetic Replacements for Human Survey Data? The Perils of Large Language Models. *Political Analysis*, 32(4):401–416, 2024.

Blair, C., Procaccia, A. D., and Tambe, M. Embeddings for Preferences, Not Semantics. 2026.

Boehmer, N., Fish, S., and Procaccia, A. D. Generative social choice: The next generation. In *Proceedings of the 42nd International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=E1E6T7KHlR.

Chen, X., Fain, B., Lyu, L., and Munagala, K. Proportionally fair clustering. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, 2019.

Cheng, M., Lee, C., Khadpe, P., Yu, S., Han, D., and Jurafsky, D. Sycophantic AI decreases prosocial intentions and promotes dependence. *Science*, 391(6792):eaec8352, 2026. doi: 10.1126/science.aec8352. URL https://www.science.org/doi/abs/10.1126/science.aec8352.

Conitzer, V., Freedman, R., Heitzig, J., Holliday, W. H., Jacobs, B. M., Lambert, N., Mosse, M., Pacuit, E., Russell, S., Schoelkopf, H., Tewolde, E., and Zwicker, W. S. Position: Social Choice Should Guide AI Alignment in Dealing with Diverse Human Feedback. In Salakhutdinov, R., Kolter, Z., Heller, K., Weller, A., Oliver, N., Scarlett, J., and Berkenkamp, F. (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 9346–9360. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/conitzer24a.html.

Coombs, C. H. *A Theory of Data*. Wiley, New York, 1964.

Cui, Z., Li, N., and Zhou, H. A large-scale replication of scenario-based experiments in psychology and management using large language models. *Nature Computational Science*, 5(8):627–634, 2025. doi: 10.1038/s43588-025-00840-7. URL https://doi.org/10.1038/s43588-025-00840-7.

DeSarbo, W. S. and Hoffman, D. L. Constructing MDS joint spaces from binary choice data: A multidimensional unfolding threshold model for marketing research. *Journal of Marketing Research*, 24(1):40–54, February 1987. doi: 10.1177/002224378702400104.

Feng, S., Sorensen, T., Liu, Y., Fisher, J., Park, C. Y., Choi, Y., and Tsvetkov, Y. Modular Pluralism: Pluralistic Alignment via Multi-LLM Collaboration. In Al-Onaizan, Y., Bansal, M., and Chen, Y.-N. (eds.), *Proceedings of the 2024 Conference on Empirical Methods 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/2024.emnlp-main.240/.

Fish, S., Gölz, 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://openreview.net/forum?id=H72JEXAPwo.

Freedman, R. Adaptive Pluralistic Alignment: A pipeline for dynamic artificial democracy, 2026. URL https://arxiv.org/abs/2605.01642.

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 Processing Systems*, volume 37, pp. 80439–80465. Curran Associates, Inc., 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/9328208f88ec69420031647e6ff97727-Paper-Conference.pdf.

Gölz, P., Haghtalab, N., and Yang, K. Distortion of ai alignment: Does preference optimization optimize for preferences? In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

He, Y., Vollen, J., and Elkind, E. Check, Please: Verifiably Fair Clustering, 2026. URL https://arxiv.org/abs/2605.12317.

Hewitt, L., Ashokkumar, A., Ghezae, I., and Willer, R. Predicting Results of Social Science Experiments using Large Language Models. *Preprint*, 2024.

Hollender, A. and Kraiczy, S. Enforcing axioms for AI alignment under loss-based rules. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=MpYSoTK65s.

Kellerhals, L. and Peters, J. Proportional fairness in clustering: A social choice perspective. In Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J., and Zhang, C. (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 111299–111317. Curran Associates, Inc., 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/c981fd12b1d5703f19bd8289da9fc996-Paper-Conference.pdf.

Kolluri, A., Wu, S., Park, J. S., and Bernstein, M. S. Fine-tuning LLMs for Human Behavior Prediction in Social Science Experiments. In Christodoulopoulos, C.,

Chakraborty, T., Rose, C., and Peng, V. (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 30096–30111, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.1530. URL https://aclanthology.org/2025.emnlp-main.1530/.

Konya, A., Schirch, L., Irwin, C., and Ovadya, A. Democratic policy development using collective dialogues and AI. *arXiv preprint arXiv:2311.02242*, 2023.

Konya, A., Thorburn, L., Almasri, W., Leshem, O. A., Procaccia, A., Schirch, L., and Bakker, M. Using collective dialogues and AI to find common ground between Israeli 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.

Maura-Rivero, R.-R., Lanctot, M., Visin, F., and Larson, K. Jackpot! Alignment as a Maximal Lottery, 2025. URL https://arxiv.org/abs/2501.19266.

Milli, S., Berker, R. E., Kraiczy, S., Shi, C., Kussman, J., Bose, A., Elkind, E., Bhattacharjee, H., Procaccia, A., and Nickel, M. For Questions of Ought, AI Could Use Some SAGE Advice. 2026.

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 Dialogue: Training and Evaluating User Language Models. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=ykSmkVqzn4.

Ni, J., Hernandez Abrego, G., Constant, N., Ma, J., Hall, K., Cer, D., and Yang, Y. Sentence-T5: Scalable Sentence Encoders from Pre-trained Text-to-Text Models. In Muresan, S., Nakov, P., and Villavicencio, A. (eds.), *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 1864–1874, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.146. URL https://aclanthology.org/2022.findings-acl.146/.

Peters, D., Pierczynski, G., and Skowron, P. Proportional participatory budgeting with additive utilities. In *Advances in Neural Information Processing Systems 34*, pp. 12726–12737, 2021.

Poole, K. T. and Rosenthal, H. A spatial model for legislative roll call analysis. *American Journal of Political Science*, 29(2):357–384, 1985. doi: 10.2307/2111172.

Robbins, H. A remark on Stirling's formula. *The American Mathematical Monthly*, 62(1):26–29, 1955. doi: 10.2307/2308012. URL https://doi.org/10.2307/2308012.

Serfling, R. J. Probability inequalities for the sum in sampling without replacement. *Annals of Statistics*, 2(1): 39–48, 1974. doi: 10.1214/aos/1176342611.

Sharma, M., McCain, M., Douglas, R., and Duvenaud, D. Who's in Charge? Disempowerment Patterns in Real-World LLM Usage. *arXiv preprint arXiv:2601.19062*, 2026.

Sorensen, T., Moore, J., Fisher, J., Gordon, M., Mireshghallah, N., Rytting, C. M., Ye, A., Jiang, L., Lu, X., Dziri, N., Althoff, T., and Choi, Y. Position: A Roadmap to Pluralistic Alignment. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

Wang, A., Morgenstern, J., and Dickerson, J. P. Large language models that replace human participants can harmfully misportray and flatten identity groups. *Nature Machine Intelligence*, pp. 1–12, 2025.

Wu, S., Choi, E., Khatua, A., Wang, Z., He-Yueya, J., Weerasooriya, T. C., Wei, W., Yang, D., Leskovec, J., and Zou, J. Humanlm: Simulating users with state alignment beats response imitation. 2026. URL https://humanlm.stanford.edu/.

Yao, S., Shinn, N., Razavi, P., and Narasimhan, K. R. τ-bench: A benchmark for tool-agent-user interaction in real-world domains. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=roNSXZpUDN.

A. Additional Details for Section 3

A.1. Counterexample for k-means

While intuitive, PRF is not guaranteed by standard clustering techniques such as, for example, k-means or even balanced k-means, as the next example demonstrates.

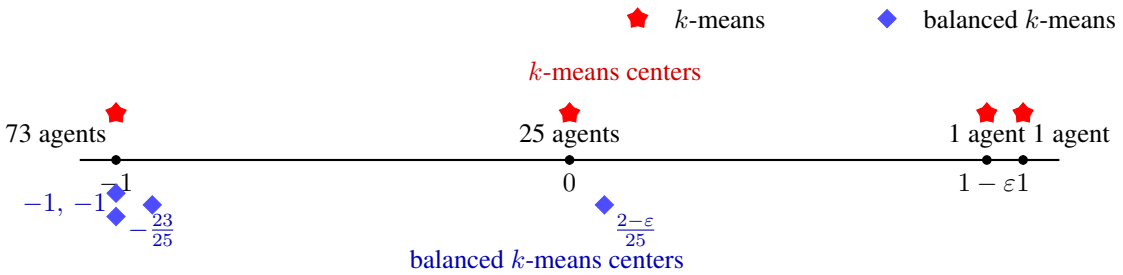

Figure A.1. The viewpoint embedding space from Example A.1, and outputs of (balanced) k-means for $k = 4$.

Example A.1. Suppose for simplicity that the viewpoint embedding space is the line segment $[-1, 1]$. There are $n = 100$ 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 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.2. Spatial Expanding Approvals Rule

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

Theorem A.1. *(Aziz et al., 2024) SEAR terminates in polynomial time $O(n^4k^2)$.*

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. This implementation removes the unnecessarily large dependency on n of n^4 by doing smart book-keeping, but additionally we are able to remove the dependency on k entirely.

Theorem 1. *SEAR can be implemented with amortized running time $O(n^2 \log n)$.*

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

Algorithm 1 Spatial Expanding Approvals Rule (SEAR)

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

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

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

    $r_p \leftarrow \Delta(p, L_p[i(p)])$
  end for
  Choose $p^* \in \arg\min_{p \in \mathcal{A}} r_p$, breaking ties by label order
  Append v_{p^*} to X
  $\rho \leftarrow n/k$ {remaining amount to collect from the selected ball}
  for $j = 1, \ldots, i(p^*)$ **do**
    $q \leftarrow L_{p^*}[j]$
    $x_q \leftarrow \min\{b_q, \rho\}$ {amount paid by agent q}
    $b_q \leftarrow b_q - x_q$ {update q's remaining budget}
    $\rho \leftarrow \rho - x_q$ {update remaining joint payment}
    if $\rho = 0$ **then**
      break
    end if
  end for
end for
return X

First compute all pairwise distances and, for every $p \in \mathcal{N}$, sort the labels $\{1, \ldots, n\}$ 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 $b_1, \ldots, b_n$. 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

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

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

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

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

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

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. Proofs Omitted from Section 4

Theorem 2. *Generate-then-Select returns a slate Y that satisfies IT-PRF.*

Proof. Let Y be the slate returned by Generate-then-select, let $R \subseteq \mathcal{N}$ be the selected corresponding humans, and write $o_j = A_j(p)$, $c_j = \psi(o_j)$, and $v_i = \phi_i(p)$. Since all embeddings lie on the unit sphere,

$$r_i(p, o) = v_i^\top \psi(o) = 1 - \frac{1}{2}\Delta(v_i, \psi(o))^2.$$

Thus, for any threshold τ, there is a radius $\rho_\tau = \sqrt{2(1-\tau)}$ such that $r_i(p, o) \geq \tau$ if and only if $\Delta(v_i, \psi(o)) \leq \rho_\tau$.

Fix an ℓ-large and τ-cohesive group $S \subseteq \mathcal{N}$. By τ-cohesiveness, for every $i, j \in S$, $r_i(p, o_j) \geq \tau$, and hence every candidate c_j with $j \in S$ lies within distance ρ_τ of every voter v_i with $i \in S$.

Suppose, for contradiction, that Y does not satisfy the IT-PRF condition for S. Then fewer than ℓ selected responses $o \in Y$ satisfy $\max_{i \in S} r_i(p, o) \geq \tau$. In particular, fewer than ℓ selected indices lie in S, since every selected o_j with $j \in S$ satisfies $\max_{i \in S} r_i(p, o_j) \geq \tau$. Because S is ℓ-large, there is some $j \in S \setminus R$.

The unselected candidate c_j is within distance ρ_τ of every voter in S. By the distance-based mPJR+ (He et al., 2026) guarantee of SEAR, the selected slate must therefore contain at least ℓ candidates x with $\min_{i \in S} \Delta(v_i, x) \leq \rho_\tau$. Equivalently, the returned response slate contains at least ℓ responses $o \in Y$ such that $\max_{i \in S} r_i(p, o) \geq \tau$, contradicting the assumption. Therefore Y satisfies IT-PRF. □

Theorem 3 (Exact Selection–Generation Commutation)**.** *Fix a prompt p, and suppose every simulator A_i, $i \in \mathcal{N}$, is faithful on p. The slate Y returned by Select-then-Generate satisfies IT-PRF.*

Proof. The claim simply follows from the fact that $f : \mathbb{S}^{d-1} \to \mathbb{R}$ defined by $f(x) = \phi_i(p)^\top x$ is maximized precisely when $x = \phi_i(p)$ since $\phi_i(p) \in \mathbb{S}^{d-1}$. Thus, since human i's viewpoint o_i^* is reward maximizing, it must be that $\psi(o_i^*) = \phi_i(p)$. In particular, if all A_i are faithful, it follows that Generate-then-select and Select-then-Generate are equivalent. □

C. Proofs Omitted from Section 5

Theorem 4 (Approximate Selection–Generation Commutation)**.** *Fix a prompt p, and suppose every simulator has fidelity gap at most α on p. Let Y be the slate returned by Select-then-Generate. Then Y satisfies γ_β-approximate IT-PRF for $\gamma_\beta = \alpha + \beta + \min\{\alpha, \beta\}$.*

Proof. For unit vectors x, y, write $\angle(x, y)$ for their angular distance. Since $x^\top y = \cos(\angle(x, y))$, the condition $r_i(p, o) \geq \cos(\theta)$ is equivalent to $\angle(\phi_i(p), \psi(o)) \leq \theta$. Thus $r_i(p, A_i(p)) \geq \cos(\alpha)$ implies $\angle(\phi_i(p), \psi(A_i(p))) \leq \alpha$ for every $i \in \mathcal{N}$.

Since $\psi(o_j)$ is within angle α of $\phi_j(p)$ for all $j \in \mathcal{N}$, the angular triangle inequality gives $\angle(\phi_i(p), \phi_j(p)) \leq \alpha + \beta$ for every $i, j \in S$. By PRF of the selected embeddings $\{\phi_i(p)\}_{i \in R}$, there are at least ℓ selected simulants $r \in R$ such that, for each such r, $\phi_r(p)$ is within angle $\alpha + \beta$ of $\phi_i(p)$ for some $i \in S$. After generation, $\psi(A_r(p))$ is within angle α of $\phi_r(p)$, so these responses are each within angle $2\alpha + \beta$ of some member of S. Equivalently, they receive reward at least $\cos(2\alpha + \beta)$ from some $i \in S$.

There is also a second, sometimes better, bound. Since by cohesiveness of S. all $\phi_i(p)$ for $i \in S$ are within angle β of the same point $\psi(o_j)$, $j \in S$, the set $\{\phi_i(p)\}_{i \in S}$ has angular diameter at most 2β. Applying PRF directly to this group, there are at least ℓ selected simulants $r \in R$ such that $\phi_r(p)$ is within angle 2β of $\phi_i(p)$ for some $i \in S$. Again using $\angle(\phi_r(p), \psi(A_r(p))) \leq \alpha$, the corresponding generated responses are within angle $\alpha + 2\beta$ of some member of S, and hence receive reward at least $\cos(\alpha + 2\beta)$.

Combining the two bounds, Y contains at least ℓ responses o' such that, for some $i \in S$,

$$r_i(p, o') \geq \cos\big(\alpha + \beta + \min\{\alpha, \beta\}\big).$$

□

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

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

Proof. For $W \in \mathcal{W}$, let $p(W) = |W|/|\mathcal{M}|$ and $\hat{p}(W) = |W \cap \mathcal{N}|/|\mathcal{N}|$. By Hoeffding's comparison theorem for sampling without replacement, the usual i.i.d. VC inequality also upper-bounds the uniform deviation probability for the present sampling model. Hence, for a sample of size n,

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

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

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

Thus, it suffices that

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

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

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

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

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

$$\log(2Ce) + 2\log(1/\varepsilon) + \log(M + L/v) \le \log(2Ce) + 3M + \frac{L}{vM} \tag{1}$$

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

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

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

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

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

□

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 induced set system on the population. We will apply Lemma C.1 to such induced set systems. The next proposition specializes this to Euclidean balls, whose induced set system has VC dimension at most $d+1$; the same argument applies to any range family $\mathcal{R}$ with VC dimension h, replacing d in the sample bound by h.

Proposition 1. *Let $\mathcal{N} \subseteq \mathcal{M}$ be a uniformly random n-element subset, where $n = \Omega(\varepsilon^{-2}(d\log(1/\varepsilon) + \log(1/\delta)))$. If X^* is a PRF slate with respect to $\{v_i\}_{i\in\mathcal{N}}$, then w.p. at least $1-\delta$, for every t-cohesive cluster S of size at least $(\ell/k+\varepsilon)m$, $\ell \in [k]$, $|\{x \in X^* \mid \min_{i\in S} \Delta(x, v_i) \le t\}| \ge \ell$.*

Proof. Let $\mathcal{N}$ be drawn uniformly without replacement from $\mathcal{P}$. 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 $\mathrm{VCdim}(\mathcal{R}) = d+1$, the same VC bound applies to $\mathcal{F}$. Thus, by the finite-population VC bound Lemma C.1, 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 $1-\delta$,

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

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

Lemma 1. *Let $\mathcal{N} \subseteq \mathcal{M}$ be a uniformly random n-element subset, where $n = \lceil \varepsilon^{-2}\log(km/\delta)\rceil$. If X^* satisfies PRF with respect to $\{v_i\}_{i\in\mathcal{N}}$, then w.p. at least $1-\delta$, X^* satisfies $(\varepsilon, 2)$-PRF[9] with respect to $\{v_i\}_{i\in M}$.*

Proof. Consider km sets defined as follows:

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

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

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

$$\Pr\big(|\overline{Y}_n - \mu| \ge t\big) \le 2\exp\left(-\frac{2n\,t^2}{\big(1-\frac{n-1}{m}\big)(B-A)^2}\right).$$

Using a Serfling-Hoeffding bound,

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

[9] Here, $(\varepsilon, 2)$-PRF means that every group $S \subseteq \mathcal{M}$ with $|S| \ge (\ell/k+\varepsilon)m$ and $D(S) \le t$ receives at least ℓ slate points within distance $2t$ of some member. The proof shows the stronger intermediate claim that some $i \in S\cap\mathcal{N}$ has ℓ slate points within distance $2t$.

Union bounding over all km sets S_v^ℓ for $v \in \mathcal{M}$ and $\ell \in [k]$, we get that

$$\Pr\left(\max_{i\in\mathcal{N},\ell\in[k]} \frac{\ell}{k} - \hat{p}_i^\ell > 0\right) \;\le\; 2km\,\exp\!\left(-2n\varepsilon^2\right),$$

thus taking $n \ge \frac{1}{2\varepsilon^2}\log\left(\frac{2km}{\delta}\right)$ suffices so that w.p. at least $1-\delta$, for all $i \in \mathcal{M}$ and $\ell \in [k]$, $\hat{p}_i^\ell \ge \frac{\ell}{k}$.

Consider any t-cohesive set S of agents of size at least $(\frac{\ell}{k}+\varepsilon)m$ where $\ell \in [k]$ and pick an arbitrary agent $v \in S$. Group S contains at least $\lceil(\ell/k+\varepsilon)m\rceil$ agents, includes v, and has diameter at most t. Thus by minimality, the group S_v^ℓ has $\mathrm{diam}(S_v^\ell) \le t$, and hence $\mathrm{diam}({S'}_v^\ell) \le t$.

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

Theorem 5. *Let $\mathcal{N} \subseteq \mathcal{M}$ be a uniformly random n-element subset, where $n = \lceil\varepsilon^{-2}\log(km/\delta)\rceil$. Fix a prompt p, and suppose every simulator A_i, $i \in \mathcal{N}$, has fidelity gap at most α on p. Let Y be the slate returned by Select-then-Generate. Then, w.p. at least $1-\delta$, Y satisfies $(\varepsilon, \gamma_\beta)$-approximate IT-PRF with respect to $\mathcal{M}$, where $\gamma_\beta = \alpha + 2\beta + 2\min\{\alpha,\beta\}$.*

Proof. For unit vectors x, y, write $\angle(x,y)$ for their angular distance. Since $x^\top y = \cos(\angle(x,y))$, the condition $r_i(p,o) \ge \cos(\theta)$ is equivalent to $\angle(\phi_i(p), \psi(o)) \le \theta$. Thus $r_i(p, A_i(p)) \ge \cos(\alpha)$ implies $\angle(\phi_i(p), \psi(A_i(p))) \le \alpha$ for every $i \in \mathcal{N}$.

Since $\psi(o_j)$ is within angle α of $\phi_j(p)$ for $j \in \mathcal{N}$, the angular triangle inequality gives $\angle(\phi_i(p), \phi_j(p)) \le \alpha + \beta$ for every $i, j \in S$. If $\mathcal{N}$ is a uniformly random subset of $\mathcal{M}$ of size $n = \lceil\varepsilon^{-2}\log(km/\delta)\rceil$, then by Lemma 1 Conditioning on the $1-\delta$ probability event of Lemma 1, we know that for any t-cohesive group S' of size at least $(\ell/k+\varepsilon)n$, there are at least ℓ points in $\{\phi_i(p) \mid i \in R\}$ that each are within distance $2t$ of some member of S'. Equivalently, at least ℓ points in $\{\phi_i(p) \mid i \in R\}$ that each are within angle 2β of some member of S, whenever $\angle(\phi_i(p), \phi_j(p)) \le \beta$ for all $i, j \in S'$. Applying this here, we get that there are at least ℓ selected simulants $r \in R$ such that, for each such r, $\phi_r(p)$ is within angle $2(\alpha+\beta)$ of $\phi_i(p)$ for some $i \in S$. After generation, $\psi(A_r(p))$ is within angle α of $\phi_r(p)$, so these responses are within angle $3\alpha + 2\beta$ of some member of S. Equivalently, they receive reward at least $\cos(3\alpha + 2\beta)$ from some $i \in S$.

There is also a second, sometimes better, bound. Since all $\phi_i(p)$ for $i \in S$ are within angle β of the same point $\psi(o_j)$ for $j \in S$, the set $\{\phi_i(p)\}_{i\in S}$ has angular diameter at most 2β. Applying the $1-\delta$ approximate PRF guarantee from above directly to this group, there are at least ℓ selected simulants $r \in R$ such that $\phi_r(p)$ is within angle 4β of $\phi_i(p)$ for some $i \in S$. Again using $\angle(\phi_r(p), \psi(A_r(p))) \le \alpha$, the corresponding generated responses are within angle $\alpha + 4\beta$ of some member of S, and hence receive reward at least $\cos(\alpha + 4\beta')$.

Combining the two bounds, Y contains at least ℓ responses o' such that, for some $i \in S$,

$$r_i(p, o') \ge \cos\!\big(\alpha + 2\beta + 2\min\{\alpha,\beta\}\big).$$

□

D. Additional Sampling Results

This section gives two additional population-transfer results for PRF under uniform sampling without replacement. First, we prove that, in Euclidean dimension d, PRF guarantees can be transferred from a sample to the full population with sample complexity depending on d, ε, and δ, but not on the population size M or the slate size k. Second, we show that the sampled-population guarantee from Lemma 1 extends to stratified sampling via a black-box conditioning argument.

As an alternative to Lemma 1, we can show a population transfer result for PRF that depends on d, similar to Proposition 1 and is independent of k and m.

Theorem D.1. *Let $\mathcal{M} \subset \mathbb{R}^d$ be a population of size m, and let k be the slate size. Let $\varepsilon, \delta > 0$, and let $\mathcal{N} \subseteq \mathcal{M}$ be a uniformly random n-element subset, where*

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

If X^* 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 $S \subseteq \mathcal{M}$, and every $t > 0$ such that $|S| \geq (\ell/k + \varepsilon)m$ and S is t-cohesive,

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

Proof. By the assumed lower bound on $|\mathcal{N}|$, Proposition 1 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

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

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 Proposition 1, there are at least ℓ members $x \in X^*$ 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 $\mathrm{conv}(A)$. 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

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

Thus at least ℓ members of X^* are within the claimed distance of some member of S. □

We next extend Lemma 1 to stratified sampling, as used in our experiments in Section 6. The proof is a black-box conditioning argument: any high-probability guarantee for a uniformly random sample can be converted into the corresponding guarantee for a uniformly random stratified sample, at the cost of a polynomial factor in the failure probability.

Suppose for $t \geq 1$, a constant, we have a partition $(P_1, \ldots, P_t)$ 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 D.1. 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 D.2. *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 $X^\star$ is a PRF slate with respect to S, then $X^\star$ 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 Lemma 1 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 $(P_1, \ldots, P_t)$). 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

$$\mathcal{S} = \Big\{\mathbf{c} = (c_1, \ldots, c_t) \in \mathbb{Z}^t_{\geq 0} : \sum_i c_i = s,\ 0 \leq c_i \leq n_i\Big\}.$$

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 replaceme) pmfs on $\mathcal{S}$:

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

We claim that

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

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

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

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

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

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

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

Proof. Recall

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

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 (†). Plugging in and cancelling the common $(m/e)^m$-terms, we obtain

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

where the error exponent is

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

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:

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

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

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

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

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

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

Therefore,

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

and hence

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

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,

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

Thus

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

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

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

This is precisely the claimed bound with

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

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. □

We thus get the direct lower bound

$$\Pr_{\mathrm{WOR}}(A) \;=\; \sum_{\mathbf{c}\in\mathcal{C}} H(\mathbf{c}) \;\geq\; H(\mathbf{c}^\star) \;\geq\; Z_t\, s^{-\frac{t-1}{2}}. \tag{†}$$

Let $\delta' > 0$ be a target failure probability for Lemma 1 under WOR sampling. Applying Lemma 1 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 $\Pr_{\mathrm{WOR}}(B) \leq \delta'$.

Set $\delta' = \delta \cdot \Pr(A)$ and use (†) 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,

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

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. □

E. Computational Complexity of Verifying PRF

Given a possible outcome (either output by our algorithms or one of the approaches we against in Section 6), we would like to check if PRF is violated by this outcome. Without any restrictions (on the metric space), this problem is computationally hard.

Theorem E.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 $V = \{v_1, v_2, \ldots, v_n\}$. 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

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

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

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

$$\Delta(c_i, c_j) \leq 2(1-\varepsilon)^2 + (n-2)\varepsilon^2 = 2 - 4\varepsilon + n\varepsilon^2. \tag{3}$$

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

$$\Delta(c_i, c_j) \geq ((c_i)_i - (c_j)_i)^2 + ((c_i)_j - (c_j)_j)^2 = 2. \tag{4}$$

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

$$2 - 4\varepsilon + 2n\varepsilon^2 = 2 - \frac{4}{n} + \frac{2n}{n^2} = 2 - \frac{2}{n} < 2. \tag{5}$$

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 $|S| = p = \frac{|\mathcal{N}|}{k}$. Then by (3), we have $\Delta(c_i, c_j) \leq 2 - 4\varepsilon + n\varepsilon^2$ for all $c_i, c_j \in S$. By (2), 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 $|S| \geq \frac{|\mathcal{N}|\ell}{k} \geq \frac{|\mathcal{N}|}{k} = p$. By (2), 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 (5). By (4), 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. □

F. Experimental Details

F.1. mPJR+

For our experiments, we evaluate satisfaction of mPJR+ (He et al., 2026).

Cohesiveness. In IT-PRF (Definition 2), the cohesiveness of a group $S \subseteq \mathcal{N}$ is defined as the minimum pairwise reward that members assign to each other's simulated responses:

$$\tau(S) := \min_{i,j \in S} r_i(p, A_j(p)).$$

Here we define the inference time version of mPJR+. In mPJR+, cohesiveness is instead defined with respect to an unselected candidate $c \in C$. In our setting, the candidate set C corresponds to the set of simulated responses $\{A_i(p)\}_{i \in \mathcal{N}}$. Overloading notation, we write:

$$\tau(S; c) := \min_{i \in S} r_i(p, c).$$

Definition F.1 (IT-mPJR+)**.** Given a prompt p, a slate of responses $X \subseteq C$ satisfies mPJR+ if for every integer $\ell \in [k]$, every unselected response $c \in C \setminus X$, every group $S \subseteq \mathcal{N}$ with $|S| \geq \ell \cdot q$, it holds that

$$\left| \left\{ o \in X \;\middle|\; \max_{i \in S} r_i(p, o) \geq \tau(S; c) \right\} \right| \geq \ell.$$

F.2. Topic Generation Prompt

We prompted Claude Opus 4.6[10] with the questions that were asked to participants in the Remesh campus protests session, and asked it to generate ten diverse topics that could be predictable from answers to these questions:

> 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?

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

> Given a person's answer to these questions, what are ten other diverse topics that 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.

F.3. Prompt Generation Prompt

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

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

F.4. Generated Topics

The following ten topics were generated by Claude Opus 4.6:

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

F.5. Simulant Prompt

Each simulant is implemented via a roleplay prompt to Llama-3.1-8B-Instruct (Grattafiori et al., 2024)[11] with the following system and user prompts.

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

System prompt.

> You are roleplaying as the following person:
>
> *{agent description}*
>
> Generate a genuine response that reflects this person's personal perspective. Do not hedge or speak in generalities — respond as this specific person would.

The participant description is constructed from the participant's demographic attributes in the Remesh survey (e.g., gender, age, political affiliation, religious affiliation, education level, urbanicity, household income).

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

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

   Here are questions you answered previously:
   Question: *{question text}*
   Response: *{participant's response}*
   [...repeated for each prior question...]

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:

F.6. Controlling the fidelity γ

To isolate the impact that fidelity γ has on the level of representation attained by different methods, we control γ semi-synthetically via constructing a human reward function $r_i^{(\gamma)}(p, o) = \phi_i^{(\gamma)}(p)^\top \psi(o)$ such that the reward that human i receives from their simulant's response $A_i(p)$ is $r_i^{(\gamma)}(p, A_i(p)) = \gamma$. We then evaluate how the level of representation attained by the methods we test changes as a function of γ.

For the response embedding ψ, we use Blair et al. (2026)'s embedding model which was trained to capture preferences on the Remesh campus protests session. Specifically, their model is trained so that distances in ψ-space predict participants' preferences rather than merely reflecting surface-level similarity: a participant i is more likely to agree with responses that are closer to their own embedded response. The accuracy of Blair et al. (2026)'s fine-tuned preference embedding model at predicting the human participants' held-out binary agreement/disagreement votes is 77.7%, compared to 63.7% from the base model (`sentence-T5-XL` (Ni et al., 2022)). We ℓ_2-normalize the outputs of Blair et al. (2026)'s embedding model, so the output of $\psi(\cdot)$ is always unit-norm.

For a target fidelity level $\gamma \in [0, 1]$, we construct the prompt-preference embedding

$$\phi_i^{(\gamma)}(p) = \gamma \cdot \psi(A_i(p)) + \sqrt{1 - \gamma^2} \cdot u_i(p),$$

where $u_i(p)$ is a unit vector orthogonal to $\psi(A_i(p))$. Thus,

$$r_i^{(\gamma)}(p, A_i(p)) = \phi_i^{(\gamma)}(p)^T \psi(A_i(p)) = \gamma\,.$$

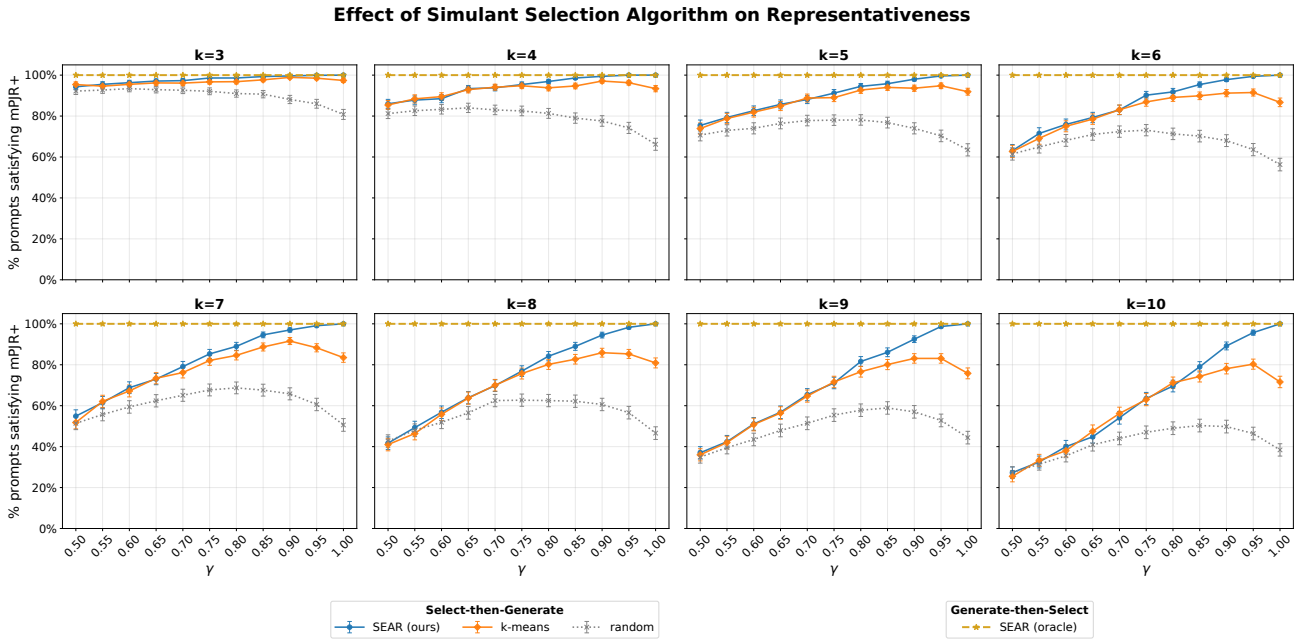

Figure F.1. **mPJR+ satisfaction rate as a function of fidelity γ and the number of selected simulants k.**

The perturbation direction u_i is generated by combining a structured global rotation direction with an idiosyncratic random direction. The structured component is obtained by applying the same block-diagonal quarter-turn operator J to every embedding,

$$(x_1, x_2, x_3, x_4, \ldots) \mapsto (-x_2, x_1, -x_4, x_3, \ldots),$$

then projecting off $\psi(A_i(p))$ and normalizing. The random component is sampled from a standard Gaussian distribution, projected to be orthogonal to both $\psi(A_i(p))$ and the structured component, and normalized. The final residual direction is

$$u_i = \rho\, s_i + \sqrt{1-\rho^2}\, r_i, \qquad \rho = 0.5,$$

where s_i is the normalized structured direction and r_i is the normalized random direction. The reason for using a structured perturbation as well as an idiosyncratic one is that adding only independent noise to each embedding causes group cohesiveness to deteriorate, resulting in all participants liking all candidates at similar levels. This makes mPJR+ trivially satisfied, even by random selection, at low levels of γ. Including both the structured and idiosyncratic components thus allows us to investigate fidelity without trivializing the mPJR+ criterion.