Diversifying Multiple Generative Agents by Aligning with Human Populations

Manh Hung Nguyen [1] **Sebastian Tschiatschek** [2] **Adish Singla** [1]

Abstract

Large Language Models (LLMs) are increasingly used in settings where systems must reflect diverse human preferences and behaviors, ranging from assistants for everyday tasks to tools for simulating human behavior in scientific research. However, prior work shows that LLMs often produce homogeneous outputs that fail to capture pluralistic human perspectives and behaviors. Rather than trying to capture this diversity using a single generative agent, we propose a framework for constructing a set of generative agents that collectively serve as faithful behavioral surrogates for a human population. Each agent is an LLM grounded in real human demonstrations (task–response pairs), whose behavior is steered to match that of a specific subpopulation. The challenge is therefore to select a representative set of agents from the exponentially large space of possible agents. We formalize this as minimizing the representation error (the average behavioral distance between each human and their nearest agent) and show that this objective is submodular, enabling methods with approximation guarantees. Extensive experiments in educational and crowdsourcing domains demonstrate that our methods construct agent sets that more faithfully represent human populations than existing methods, and that these agents reproduce subpopulation-level behavioral patterns on unseen tasks.

1. Introduction

Generative AI is evolving at a remarkable pace, leading to a growing deployment across many domains. These applications span from serving as personal assistants for everyday tasks to functioning as tools for simulating human behavior in scientific research. However, prior work has shown that these models tend to produce homogeneous outputs and converge on a narrow set of responses (Bao et al., 2024; Wenger & Kenett, 2025; Lee et al., 2024; Lahoti et al., 2023; Wu et al., 2025; Goel et al., 2025; Shumailov et al., 2024; Kirk et al., 2024). This tendency reflects an "Artificial Hivemind" effect (Jiang et al., 2025), in which LLMs produce homogeneous responses both within a single model and across multiple models. Such homogenization limits the applicability of generative models in domains that require rich diversity of perspectives and behaviors. Examples include NLP tasks such as paraphrasing (Cegin et al., 2023), simulating human opinions and behaviors in surveys (Xie et al., 2024; Santurkar et al., 2023), evaluating recommendation systems (Yoon et al., 2024), replicating human-subject studies (Aher et al., 2023), and simulations in educational contexts (Markel et al., 2023; Zhang et al., 2025; Nguyen et al., 2024; 2025; Scholz et al., 2025; Nguyen et al., 2026). As adoption of generative AI accelerates, such homogenization creates a central obstacle for pluralistic alignment (Sorensen et al., 2024), because AI systems should not only perform well on average, but also reflect the diversity of human populations.

Addressing these limitations requires models that can faithfully represent the diverse spectrum of human behaviors. However, standard alignment approaches, such as Reinforcement Learning from Human Feedback (RLHF) (Ouyang et al., 2022), typically rely on aggregated preferences, which can obscure the full range of human perspectives and may be unrepresentative (Casper et al., 2023; Kirk et al., 2023). Indeed, prior work suggests that it is fundamentally challenging, if not infeasible, to train one model to simultaneously satisfy a multitude of diverse, potentially conflicting, preferences (Ouyang et al., 2022). Recent work focuses on conditioning LLMs on personas or demographic attributes to capture diverse behaviors (Salewski et al., 2023; Tao et al., 2024; Aher et al., 2023). Examples include generating diverse data via attributed prompts (Yu et al., 2023), selecting effective personas (Choi & Li, 2024), or learning mixture-of-personas models (Bui et al., 2025). However, these methods typically focus on steering a single LLM agent, relying on predefined personas and attributes or requiring access to model logits and fine-tuning. Moreover, the resulting agents are not jointly optimized for population-level coverage.

In this work, we move beyond single-agent and heuristic approaches to propose a framework for the principled construction of a set of faithful behavioral surrogates for a

[1]MPI-SWS, Germany [2]University of Vienna, Austria. Correspondence to: Manh Hung Nguyen <manguyen@mpi-sws.org>.

Pluralistic Alignment Workshop @ ICML 2026, Seoul, South Korea.

human population. Our guiding hypothesis is that a curated set of agents, each grounded in real human demonstrations, can collectively achieve a more faithful behavioral representation of a human population than any single model. We achieve this by leveraging in-context learning, where each agent's behavior is steered by a small set of demonstrations. We formulate the construction of this agent set as a combinatorial optimization over their prompts. To make this problem tractable, we cast it as a submodular optimization task and propose several methods that offer different trade-offs between performance and computational complexity. Our evaluation across educational and crowdsourcing domains shows that these methods construct agents that faithfully replicate subpopulation-level behavioral patterns, including on held-out tasks not used during agent construction.

In summary, our main contributions are:

- We formalize the problem of constructing a representative set of behavioral surrogates for a human population as a submodular optimization problem. (§3).
- We propose a human-mapped proxy construction that reduces the combinatorial agent search space to linear size, along with methods offering different computational complexity and performance trade-offs (§4).
- We empirically demonstrate across educational and crowdsourcing domains that our methods preserve subpopulation-level behavioral fidelity, outperforming existing persona-based and heuristic approaches. (§5).

2. Related Work

Diversity and biases in LLM outputs. Current LLMs often generate outputs characterized by limited diversity and systematic biases, making them unrepresentative of many population groups. Studies have revealed systematic homogeneity in LLM responses (Yoon et al., 2024) and divergence from the distribution of real human survey responses (Sun et al., 2024). This phenomenon, recently termed the "Artificial Hivemind" (Jiang et al., 2025), has been widely reported across various models (Bao et al., 2024; Wu et al., 2025; Goel et al., 2025; Shumailov et al., 2024). Such limitations often come from biases present in training data (Brown et al., 2020), leading to significant misalignment with diverse groups (Santurkar et al., 2023; Tao et al., 2024). This failure to capture a spectrum of human viewpoints is further exemplified by the "hyper-accuracy distortion" observed in human subject study replications (Aher et al., 2023). Collectively, these findings underscore fundamental challenges in representing population diversity and raise concerns about the tendency of AI to perpetuate dominant community norms (Bender et al., 2021; Liu et al., 2024). Our work addresses this by constructing a set of LLM agents that collectively serve as faithful behavioral surrogates for a target population, grounded in demonstrations rather than personas or demographic metadata.

Inducing and aligning behaviors of LLMs. Recent work has explored various techniques to align LLMs with specific population groups to address representation issues. By treating LLMs as a superposition of perspectives (Kovac et al., 2023), researchers have proposed prompting methods such as in-context impersonation (Salewski et al., 2023), demographic-aware prompting (Santurkar et al., 2023), and persona-based conditioning (Yu et al., 2023; Choi & Li, 2024; Bui et al., 2025). Beyond prompting, other efforts focus on fine-tuning models for cross-cultural (Ramezani & Xu, 2023) or value alignment (Liu et al., 2022). However, standard alignment methods like RLHF (Ouyang et al., 2022) often rely on aggregated preferences, which can obscure certain viewpoints and fail to capture broad human diversity (Casper et al., 2023; Kirk et al., 2023). Our work builds on in-context learning but addresses a fundamentally different problem. Rather than steering a single agent toward one persona or attribute profile, we optimize an ensemble of agents for population-level behavioral coverage.

LLM-based crowdsourcing and behavior simulation. LLMs have increasingly been adopted to facilitate crowdsourcing and simulate human behavior across diverse domains. This shift enables research at scale while mitigating risks traditionally associated with human-subject studies. Recent work has demonstrated the effectiveness of LLMs as synthetic crowdworkers for core NLP tasks, including data annotation (Moskovskiy et al., 2024), paraphrasing (Cegin et al., 2023), information extraction (Zhang et al., 2023), and text classification (Sun et al., 2023). Beyond standard NLP, LLMs are used to simulate human behaviors and opinions (Xie et al., 2024; Santurkar et al., 2023), evaluate interactive recommendation systems (Yoon et al., 2024), and replicate human subject studies (Aher et al., 2023). In educational contexts, these models serve as simulated learners or tutors, capturing a range of student profiles to improve instructional design (Nguyen et al., 2024; Markel et al., 2023; Zhang et al., 2025). While these works demonstrate LLMs' potential as individual surrogates, they typically deploy a single model or ad-hoc persona per task and do not address how to systematically construct an agent set that covers the behavioral diversity of a human population. Our framework addresses this gap through optimization over the agent space.

3. Problem Formulation

We first introduce necessary notation and background. We then formally define the problem of selecting a representative set of agents and the optimization objective.

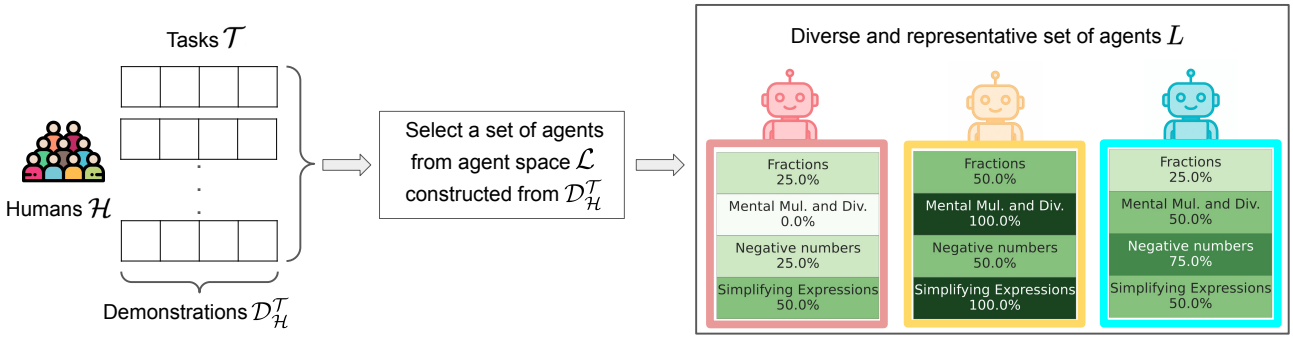


Figure 1. **Illustrative example.** In this example, $\mathcal{H}$ is a group of students with diverse skill levels working on math tasks $\mathcal{T}$. Each agent is an LLM grounded in real student demonstrations, and the selected set L is optimized to cover the behavioral space of the population. The resulting agents exhibit distinct proficiency profiles across mathematical concepts, each corresponding to a subpopulation of students matched by skill level. Crucially, these behavioral patterns transfer to held-out tasks not used during agent construction.

3.1. Preliminaries

Let $\mathcal{H} = \{h_1, h_2, \ldots, h_N\}$ denote a population of N humans. We assume access to a set of demonstrations $\mathcal{D}_{\mathcal{H}}^{\mathcal{T}}$ consisting of task–response pairs provided by humans in $\mathcal{H}$ across tasks $\mathcal{T}$. We define an agent l as an LLM whose behavior is steered through in-context learning by a prompt containing a subset of K demonstrations from $\mathcal{D}_{\mathcal{H}}^{\mathcal{T}}$. This formulation allows us to construct a large space of potential agents $\mathcal{L}$, where each agent $l \in \mathcal{L}$ is anchored by a specific combination of human demonstrations. The number of possible agents is $|\mathcal{L}| = \binom{|\mathcal{D}_{\mathcal{H}}^{\mathcal{T}}|}{K}$, which for fixed K scales as $(|\mathcal{T}| \cdot |\mathcal{H}|)^K$. This large space $\mathcal{L}$ reflects the diverse behavioral spectrum present in the population.

To quantify and compare behaviors, we represent both humans and agents as embedding vectors in $\mathbb{R}^d$ computed from their responses to the tasks in $\mathcal{T}$. Each human h is thus characterized by a vector $\mathbf{e}_h$ summarizing their behavioral patterns, while each agent l is similarly represented by a vector $\mathbf{e}_l$. In Section 5, we will specify the embedding types used for each experimental domain, which range from binary performance vectors and discrete answer profiles to continuous semantic representations.

3.2. Objective

We measure the behavioral similarity between humans and agents through the distance between their embedding vectors, denoted by $\mathrm{dist}(\mathbf{e}_1, \mathbf{e}_2)$ (e.g., Euclidean distance). To quantify how well a set of agents $L \subseteq \mathcal{L}$ represents the population $\mathcal{H}$, we define the *representation error* $g(L)$ as the average distance from each human to their closest agent: $g(L) = \frac{1}{|\mathcal{H}|} \sum_{h \in \mathcal{H}} \min_{l \in L} \mathrm{dist}(\mathbf{e}_h, \mathbf{e}_l)$. By minimizing this error, we aim to select a set L of size M that aligns with the human population and captures its inherent diversity: $L^{\text{opt}} = \arg\min_{L \subseteq \mathcal{L}, |L| \leq M} g(L)$. We define the baseline error when no agents are selected as $g(\emptyset) = \frac{1}{|\mathcal{H}|} \sum_{h \in \mathcal{H}} D_{\max}$, where $D_{\max}$ is a constant exceeding any possible distance between a human and an agent. The objective then becomes maximizing the *average distance reduction* $f(L) = g(\emptyset) - g(L)$:

$$f(L) = \frac{1}{|\mathcal{H}|} \sum_{h \in \mathcal{H}} \left[D_{\max} - \min_{l \in L} \mathrm{dist}(\mathbf{e}_h, \mathbf{e}_l) \right].$$

The optimization problem is thus to find a subset of M agents that maximizes this average distance reduction: $L^{\text{opt}} = \arg\max_{L \subseteq \mathcal{L}, |L| \leq M} f(L)$. A key property of the objective function $f(L)$ is that it naturally encourages diversity in the selected set of agents. Since embeddings are computed from actual task responses (not from demonstrations), representation error directly measures proximity in behavioral response space. In Section 5, we empirically verify that minimizing representation error on training tasks yields agents whose behaviors remain closely aligned with humans on held-out validation tasks.

4. Methodology

We establish the hardness of the representative agent selection problem and prove the submodularity of our objective function. We then discuss how naive applications of existing strategies are insufficient and present improved methods in Section 4.2. The performance guarantees and time complexity of the methods discussed in this section are in Table 1.

4.1. Hardness of the Problem and Connection to Submodularity

Proposition 4.1 (NP-Hardness)**.** *The problem of selecting an optimal subset $L^* \subseteq \mathcal{L}$ of size M that maximizes $f(L)$ is NP-hard.*

This follows from the NP-hardness of the facility location problem (Verter, 2011); we provide full proofs in Ap-

Method	Performance Guarantee	Time Complexity
OPT	$f(L^{\text{opt}})$	$\mathcal{O}\big((\lvert\mathcal{T}\rvert \cdot \lvert\mathcal{H}\rvert)^{KM}\big)$
SINGLE	–	$\mathcal{O}(1)$
RANDOM	–	$\mathcal{O}(M)$
GREEDY	$(1-\frac{1}{e}) \cdot f(L^{\text{opt}})$	$\mathcal{O}\big(M \cdot \lvert\mathcal{H}\rvert \cdot (\lvert\mathcal{T}\rvert \cdot \lvert\mathcal{H}\rvert)^{K}\big)$
SAMPLEGREEDY	–	$\mathcal{O}\big(M \cdot \lvert\mathcal{H}\rvert \cdot \psi \cdot (\lvert\mathcal{T}\rvert \cdot \lvert\mathcal{H}\rvert)^{K}\big)$
REPPOP$_{\text{demo}}$	–	$\mathcal{O}\big(M \cdot K \cdot \lvert\mathcal{T}\rvert \cdot \lvert\mathcal{H}\rvert^2\big)$
REPPOP$_{\text{mapped-1}}$	$(1-1/e) \cdot (\gamma \cdot f(L^{\text{opt}}) - \rho_1)$	$\mathcal{O}\big(K \cdot \lvert\mathcal{H}\rvert + M \cdot \lvert\mathcal{H}\rvert^2\big)$
REPPOP$_{\text{mapped-2}}$	$(1-1/e) \cdot (\gamma \cdot f(L^{\text{opt}}) - \rho_2)$	$\mathcal{O}\big(K \cdot \lvert\mathcal{T}\rvert \cdot \lvert\mathcal{H}\rvert + M \cdot \lvert\mathcal{H}\rvert^2\big)$

Table 1. Performance guarantees and time complexity analysis.

pendix E.1 for completeness. While the facility location formulation has been applied to diversity-accuracy trade-offs in recommender systems (Panteli & Boutsinas, 2023), our setting differs in that we construct generative agents from a combinatorially large space rather than re-ranking a fixed item set.

Proposition 4.2 (Submodularity of the Objective Function $f(L)$)**.** *The objective function* $f(L) = \frac{1}{|\mathcal{H}|}\sum_{h\in\mathcal{H}}[D_{\max} - \min_{l\in L}\operatorname{dist}(\mathbf{e}_h, \mathbf{e}_l)]$ *is submodular.*

This follows from the submodularity of facility location objectives (Krause & Golovin, 2014); we provide full proofs in Appendix E.2 for completeness.

Greedy Approximation. Due to the NP-hardness of the problem, finding the optimal solution L^{opt} requires exhaustive search with time complexity $\mathcal{O}((|\mathcal{T}| \cdot |\mathcal{H}|)^{KM})$. The submodularity of our objective function enables the GREEDY algorithm to achieve a $(1 - 1/e) \cdot f(L^{\text{opt}})$-approximation guarantee (Nemhauser et al., 1978), with time complexity $\mathcal{O}(M \cdot |\mathcal{H}| \cdot (|\mathcal{T}| \cdot |\mathcal{H}|)^K)$. However, this approach becomes intractable as the agent space grows. Following prior work (Singla et al., 2014; Mirzasoleiman et al., 2015), one can fix a candidate pool $\mathcal{C}$ containing only a fraction ψ of agents from $\mathcal{L}$ and apply greedy selection within this pool (SAMPLEGREEDY). At each round, the marginal contribution of each remaining agent in $\mathcal{C}$ is evaluated relative to the current set, and the best agent is added. This process continues until M agents are selected. The time complexity is $\mathcal{O}(M \cdot |\mathcal{H}| \cdot \psi \cdot (|\mathcal{T}| \cdot |\mathcal{H}|)^K)$, which is still impractical for large populations and motivates the more efficient methods introduced in the following.

4.2. Our Proposed Methods

Greedy selection of demonstrations for an agent's context. Searching through the exponentially large agent space $\mathcal{L}$ is computationally infeasible, limiting the practicality of standard greedy methods. Hence, we propose an alternative method, REPPOP$_{\text{demo}}$ (**Rep**resentative **Pop**ulation using **demo**nstration-level greedy selection), which reduces the complexity to $\mathcal{O}(M \cdot K \cdot |\mathcal{T}| \cdot |\mathcal{H}|^2)$—scaling *linearly* in both M and K, eliminating the exponential dependence on K. Instead of enumerating all candidate agents and evaluating their marginal gains, REPPOP$_{\text{demo}}$ builds each agent incrementally (cf. Algorithm 1 in Appendix D). At each step it greedily selects a demonstration from the pool $\mathcal{D}_{\mathcal{H}}^{\mathcal{T}}$ to extend the current context Ω. We denote by l_Ω the agent constructed from Ω. This demonstration-level greedy construction avoids the exponential blow-up in $|\mathcal{L}|$, but sacrifices the formal performance guarantee of standard greedy selection. While submodular methods have been used for in-context example selection to maximize single-model task performance (Ji et al., 2024; Qian et al., 2024; Trust & Minghim, 2023), we solve a multi-agent coverage problem over a combinatorially large space.

Greedy selection of human-mapped agents. To further address the computational intractability of searching the full agent space $\mathcal{L}$, we introduce a reduced pool of proxies that directly reflect the humans in the population. We construct $\tilde{\mathcal{L}} = \{l_h \mid h \in \mathcal{H}\}$, where each agent l_h corresponds to a human $h \in \mathcal{H}$ and is formed by conditioning on a subset of K demonstrations from $\mathcal{D}_h^{\mathcal{T}}$. This one-to-one mapping reduces the candidate space to $|\tilde{\mathcal{L}}| = |\mathcal{H}|$ while preserving diversity. The selection problem then becomes $L^* = \arg\max_{L\subseteq\tilde{\mathcal{L}}, |L|\le M} f(L)$.

Building on this human-centered mapping idea, we instantiate two methods, REPPOP$_{\text{mapped-1}}$ and REPPOP$_{\text{mapped-2}}$, which both follow the general procedure in Algorithm 2 (Appendix D). Their only difference lies in how demonstrations are selected to construct a human-mapped agent (line 5). In REPPOP$_{\text{mapped-1}}$, the K demonstrations for each human are sampled uniformly at random, yielding lightweight proxy agents with cost $\mathcal{O}(K|\mathcal{H}|)$. In contrast, REPPOP$_{\text{mapped-2}}$ selects the K demonstrations greedily with respect to the human's own behavior, producing stronger proxies at cost $\mathcal{O}(K|\mathcal{T}||\mathcal{H}|)$. Concretely, for each human h, the demonstrations are chosen to minimize the distance between the human's embedding $\mathbf{e}_h$ and the embedding of the constructed

Dataset	Domain	Task Type	Representation Type	Multimodal	No. Humans	No. Tasks
EEDI	Education	Multi-choice	Performance	No	50	40
OpinionQA	Crowdsourcing	Multi-choice	Opinion	No	500	77
WikiArt	Crowdsourcing	Open-ended	Semantic	Yes	100	20

Table 2. Statistics of datasets used in our experiments.

agent $\mathbf{e}_{l_h}$, i.e., $\mathrm{dist}(\mathbf{e}_h, \mathbf{e}_{l_h})$. Both methods share the same greedy selection stage over the proxy pool, which requires $\mathcal{O}(M|\mathcal{H}|^2)$ time. Crucially, both methods scale *linearly* in M and K, avoiding the exponential dependence that renders standard greedy methods impractical. Additionally, both enjoy the same approximation guarantee in Theorem 4.3.

Theorem 4.3 (Performance Guarantee for REPPOP$_{\text{mapped-1}}$ and REPPOP$_{\text{mapped-2}}$)**.** *Let $\tilde{\mathcal{L}} = \{l_h | h \in \mathcal{H}\}$ be the proxy agent set where for each $h \in \mathcal{H}$, $l_h \in N_\rho(h)$, with $N_\rho(h)$ representing the ρ-neighborhood of h. Define the human coverage ratio $\gamma = f(L^*_{\mathcal{H}})/f(L^*_{\mathcal{L}}) \in [0,1]$, where $L^*_{\mathcal{H}}$ is the optimal subset from the human set and $L^*_{\mathcal{L}}$ is the optimal subset from the full agent set. If $L^{greedy}_{\tilde{\mathcal{L}}}$ is the subset of size M returned by the greedy algorithm on $\tilde{\mathcal{L}}$, then:*

$$f(L^{greedy}_{\tilde{\mathcal{L}}}) \;\geq\; (1-1/e)\left(\gamma \cdot f(L^*_{\mathcal{L}}) - \rho\right),$$

where γ measures the cost of restricting the search space to humans (coverage quality) and ρ measures the cost of approximating each human by a proxy agent (imitation error). The value of γ is determined by how expressive the human set is relative to the full agent space, whereas ρ depends on the proxy construction strategy: uniform sampling in REPPOP$_{\textit{mapped-1}}$ *typically yields larger ρ, while greedy selection in* REPPOP$_{\textit{mapped-2}}$ *achieves smaller ρ at the expense of higher computational cost.*

The bound in Theorem 4.3 decomposes the approximation cost into two independent sources. (1) **Agent space restriction** (γ): by limiting our search to one agent per human instead of the full combinatorial space $\mathcal{L}$, we may miss some behavioral profiles. γ close to 1 indicates that the human set is expressive enough that this restriction loses little. (2) **Proxy imitation** (ρ): for each human h, the proxy agent l_h may not perfectly replicate h's behavior. ρ quantifies this per-agent imitation error, which is controlled by the proxy construction strategy (REPPOP$_{\text{mapped-1}}$ vs. REPPOP$_{\text{mapped-2}}$). In practice, REPPOP$_{\text{mapped-2}}$ targets a smaller ρ by greedily selecting demonstrations that minimize $\mathrm{dist}(\mathbf{e}_h, \mathbf{e}_{l_h})$ for each human.

Proof sketch. We show that for the optimal human subset $L^*_{\mathcal{H}}$, the corresponding set of proxy agents $L^*_{\tilde{\mathcal{H}}} = \{l_h | h \in L^*_{\mathcal{H}}\} \subseteq \tilde{\mathcal{L}}$ satisfies $f(L^*_{\tilde{\mathcal{H}}}) \geq f(L^*_{\mathcal{H}}) - \rho$ due to the ρ-neighborhood property. Since $L^*_{\tilde{\mathcal{L}}}$ is optimal within $\tilde{\mathcal{L}}$, we have $f(L^*_{\tilde{\mathcal{L}}}) \geq f(L^*_{\tilde{\mathcal{H}}})$. By the standard greedy approximation for submodular functions, $f(L^{\text{greedy}}_{\tilde{\mathcal{L}}}) \geq (1-1/e) f(L^*_{\tilde{\mathcal{L}}})$. Combining these inequalities and using the human coverage ratio $\gamma = f(L^*_{\mathcal{H}})/f(L^*_{\mathcal{L}})$, we derive our bound (cf. Appendix E.3 for proof). Regarding runtime, the linear scaling of our methods translates to substantial practical speedups, which will be shown in Section 5.

5. Experimental Evaluation

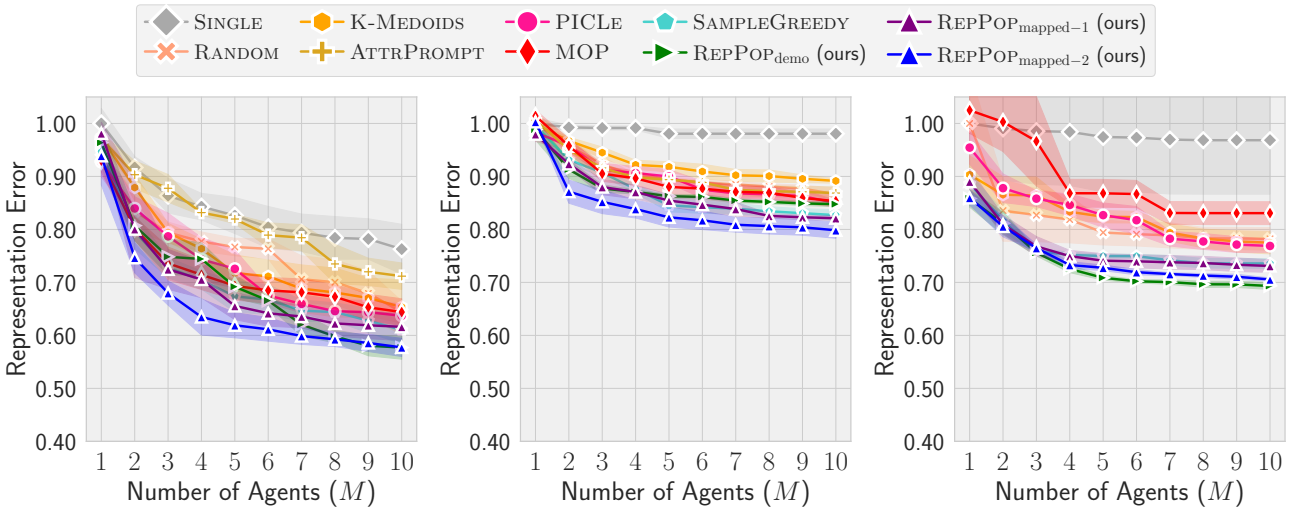
We present the evaluation domains and datasets (see Table 2 for summary statistics, and Figure 5 in Appendix B for task examples), describe the methods evaluated, and discuss the main results. We demonstrate our agents functioning as effective surrogates for human populations. Additional setup details are in Appendix B and further results in Appendix C.

5.1. Evaluation Domains and Datasets

Education: Math Questions and Answering (EEDI). LLMs capturing diverse student behavior can help teachers practice instructional strategies (Markel et al., 2023) and conduct virtual pretesting (Benedetto et al., 2024) in a safe environment. We use the EEDI dataset (Wang et al., 2020), containing multiple-choice math questions and answers collected from students. We use a subset covering 50 students and 40 exercises, split 50/50 into training and validation tasks. Each student or agent is represented by a binary embedding of correct/incorrect answers (cf. Appendix B.1 for additional details).

Crowdsourcing: Opinion Survey (OpinionQA). LLMs can serve as surrogates for survey respondents, producing answers that reflect the diverse opinions different groups of people express. We use the OpinionQA dataset (Santurkar et al., 2023), which contains multiple-choice questions from the Pew American Trends Panel (ATP W92 survey, 77 politics questions) together with human responses. Our subset covers 500 respondents and 77 questions, split into 40 training and 37 validation tasks. Each person or agent is represented by a vector of ordinal-mapped, normalized responses (cf. Appendix B.2 for additional details).

Crowdsourcing: Data Annotation (WikiArt). LLMs can serve as surrogates for annotators in tasks where diverse perspectives are valuable, such as labeling emotions evoked by art (Mohammad & Kiritchenko, 2018; Mohamed et al., 2022) (example in Figure 5c, Appendix B.3). Since existing datasets lack individual-level annotations, we use 100 LLM-



(a) EEDI *(b)* OpinionQA *(c)* Wikiart

Figure 2. **Representation Error.** We report the representation error of each method as the number of agents increases. Errors are normalized by the error of SINGLE, such that values less than 1.0 indicate improvement over the single-agent baseline. We report the means and standard errors (error bars) over three runs with different seeds. We exclude the ATTRPROMPT method from the plots in WikiArt domain as its error consistently stays around 4.0, far from the current range in the plots. Our methods consistently achieve lower representation error than the baselines across all three datasets.

based annotators (Gemma3-27B conditioned on Big Five personality profiles) as a proof-of-concept for open-ended tasks. Our subset covers 100 annotators and 20 WikiArt paintings (Tan et al., 2019), split 50/50 into training and validation tasks. Each annotator or agent is represented by a continuous LLM embedding of their responses (cf. Appendix B.3 for additional details).

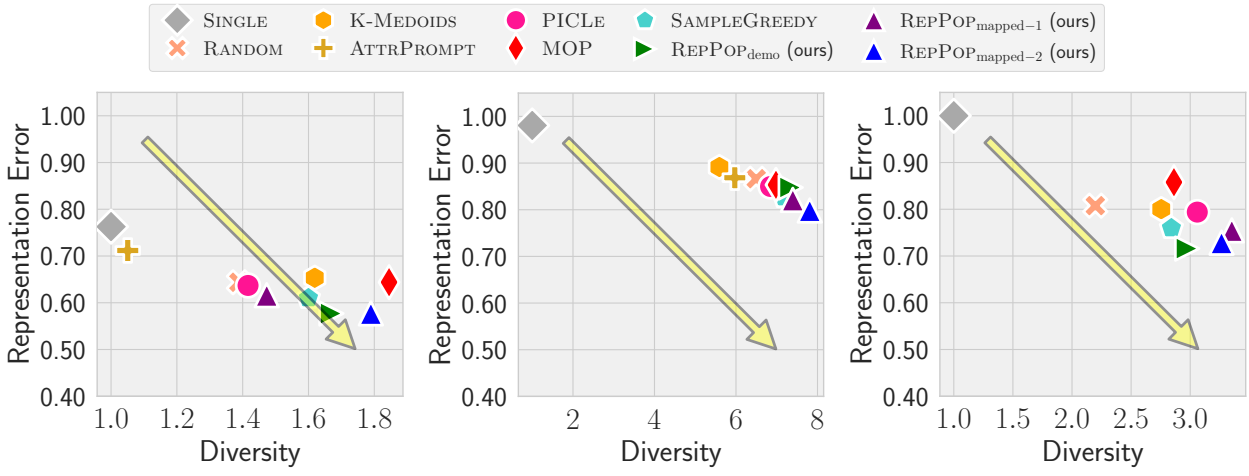
5.2. Methods Evaluated

We compare our proposed methods REPPOP$_{\text{demo}}$, REPPOP$_{\text{mapped-1}}$, and REPPOP$_{\text{mapped-2}}$ (cf. Section 4.2) against SAMPLEGREEDY (cf. Section 4.1) and the following baselines. The SINGLE baseline uniformly samples a single agent from $\mathcal{L}$ and performs M rollouts, while RANDOM baseline uniformly selects M agents from $\mathcal{L}$ and performs one rollout for each. The K-MEDOIDS baseline applies K-medoids clustering to form M clusters of humans, and for each cluster uniformly samples K demonstrations from the humans in that cluster to construct an agent; this approach requires re-clustering whenever a new agent is added. For SAMPLEGREEDY, we set the sample size to the number of humans in all experiments. For REPPOP$_{\text{demo}}$, we use a stochastic greedy variant with subsample size $\alpha = 100$ (WikiArt, EEDI) and $\alpha = 1000$ (OpinionQA). All agents use a decoding temperature of 1.0.

We also compare against recent personality/role-playing methods adapted to our setting. ATTRPROMPT (Yu et al., 2023) uses an LLM to generate attribute dimensions (e.g., knowledge level, political orientation) and values; agents are created by randomly sampling attribute combinations and generating synthetic demonstrations. PICLE (Choi & Li, 2024) samples a human, fine-tunes a model on their demonstrations, and selects the top-K demonstrations maximizing the likelihood ratio between the fine-tuned and base models. MIXTURE-OF-PERSONAS (MOP) (Bui et al., 2025) clusters demonstrations into personas and trains a gating network to weigh exemplars; we adapt it to construct M static agents by selecting the top-K exemplars that maximize the gating score for the average cluster context.

5.3. Results

Representation error. We evaluate each method by measuring the representation error on the held-out validation tasks $\mathcal{T}_{\text{val}}$. All errors are normalized by the single-agent baseline (SINGLE) at M=1, so that values below 1.0 indicate improvement (lower is better). Figure 2 shows results for varying M agents (underlying model is Gemma3-12B) with K=3. Among simple baselines, RANDOMimproves over SINGLE, but K-MEDOIDSdoes not further improve, suggesting naively clustering humans does not reduce representation error. Personality-based methods (ATTRPROMPT, PICLE, MOP) yield only modest gains, indicating that personality prompting alone is insufficient to capture the spectrum of human responses. SAMPLEGREEDY, which exploits the submodularity of the objective, emerges as the strongest baseline. Our methods achieve the best performance across all three datasets. REPPOP$_{\text{mapped-2}}$ significantly outperforms SAMPLEGREEDY($p < 0.01$, paired t-test) and generalizes



(a) EEDI *(b)* OpinionQA *(c)* WikiArt

Figure 3. **Diversity and Representation Error.** We visualize diversity (average pairwise distance between response-derived agent embeddings) and representation error for $M = 10$ agents with $K = 3$ demonstrations, both normalized by the SINGLEbaseline. The yellow arrow indicates the direction of improved performance (higher response diversity, lower error). In EEDI, while MOPshows slightly higher diversity, it incurs higher representation error, suggesting unfaithful diversity. Our methods outperform baselines on both metrics in OpinionQA and WikiArt. These results show that explicitly minimizing representation error with respect to a human population yields agents with genuinely diverse response behaviors.

to unseen tasks. We observe the same trend across context sizes K, generative models spanning 4B–70B parameters, and embedding models (cf. Appendix C).

Diversity and Representation Error. We analyze whether minimizing representation error also yields agents with diverse response behaviors. Again, we note that the embeddings used here are computed from agents' actual generated responses, not from their prompts. Figure 3 visualizes diversity (average pairwise distance between agent response embeddings) against representation error for $M = 10$, $K = 3$. Ideally, an effective method should achieve both high diversity and low representation error (bottom-right direction). Our methods have a tendency to occupy this desirable region across all three domains. In EEDI, while MOP shows slightly higher diversity, it incurs notably higher representation error, indicating that its diversity does not faithfully reflect real student behavioral patterns. By explicitly minimizing representation error, our methods recover genuine population-level diversity.

Agents as Surrogates on New Tasks. We evaluate if agents function as behavioral surrogates for human subgroups in EEDI domain as an example (cf. Appendix C.3 for analyses on other domains). Figure 4 shows that REPPOP$_{\text{mapped-2}}$ constructs agents covering diverse student profiles. For example, Agent 4 excels in Mental Multiplication but struggles with Fractions, whereas Agent 3 captures a balanced but poor skill set. When queried on new questions, the agents' performance across math concepts closely mirrors the proficiency levels of the students they represent, as shown in Figure 1. This confirms that agents constructed by our method can work as surrogates for faithfully producing population behavioral nuances.

Runtime Analysis. We analyze the runtime for constructing an agent set on EEDI with $M = 10$ (cf. Appendix B.6 for resources). Both REPPOP$_{\text{mapped-1}}$ and REPPOP$_{\text{mapped-2}}$ are computationally efficient: REPPOP$_{\text{mapped-1}}$ takes about 29s ($K = 1$) and 33s ($K = 5$), while REPPOP$_{\text{mapped-2}}$ takes about 35s ($K = 1$) and 64s ($K = 5$). These variants parallelize proxy-agent construction across the human population, so the reported time covers a single proxy build plus the greedy selection step. In contrast, REPPOP$_{\text{demo}}$ is significantly more expensive (about 355s for $K = 1$, about 1854s for $K = 5$), aligning with its $\mathcal{O}(M \cdot K \cdot |\mathcal{H}|^2)$ iterative search over the demonstration pool (cf. Table 1), whereas mapped variants reduce the search space to $|\mathcal{H}|$ pre-constructed proxies.

6. Concluding Discussions

Summary. We studied constructing representative agents that collectively serve as faithful behavioral surrogates for human populations. Casting this as a submodular optimization problem, we developed methods with a human-mapped proxy construction offering scalable trade-offs between computational complexity and approximation guarantees. Extensive experiments across educational and crowdsourcing domains demonstrated that our methods construct agents that better cover population behavioral spaces than heuristic, persona-based, and mixture-model

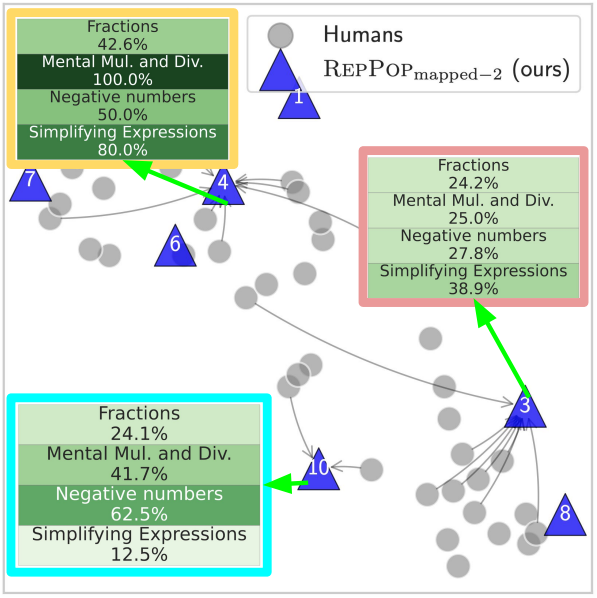


Figure 4. **Agents as Human Surrogates.** We visualize 2D embeddings of humans and agents selected by REPPOP$_{\text{mapped-2}}$ on EEDI training tasks using UMAP (McInnes & Healy, 2018). Selected agents cover diverse regions of the human population. We show aggregated metadata (in boxes) for the human groups covered by specific agents (connected by gray arrows). Each agent captures a distinct human (student) profile, representing a group with particular math concepts mastery. When queried on new exercises, they show matched skills and performances as shown in Figure 1 with same color coding of agent boxes.

baselines. Crucially, this behavioral fidelity transfers to held-out tasks, enabling the use of these agents as surrogates in downstream applications.

Limitations and future work. We discuss several limitations of our work and future work. First, we relied on prompting and in-context learning for constructing agents, future work could explore fine-tuning or retrieval-augmented approaches. Second, while our EEDI and OpinionQA experiments rely on real human data, the WikiArt task utilizes synthetic annotators (included as a proof of concept for open-ended visual tasks), which could potentially introduce circularity. Evaluating on real human annotation datasets in such domains is an important next step. Third, our coverage-based objective prioritizes behavioral breadth; extending this with proportional representation will be necessary for applications requiring population-level statistical inference. Finally, expanding our experiments to a larger scale across multiple LLM families, temperatures, demonstration orderings, and integrating advanced agent architectures that support memory and multi-turn interaction (Wu et al., 2024; Wang et al., 2024) would be interesting future work.

Ethics Statement

This work uses publicly available datasets (EEDI, OpinionQA, and WikiArt) that have been anonymized and do not contain personally identifiable information. No new human subjects were recruited, and thus no IRB approval was required. Potential risks of this work include misuse of representative agents to simulate or stereotype demographic groups. Our methods are intended solely for research and educational purposes and not for applications that could cause harm to individuals or communities. We acknowledge that language models can reflect and amplify biases present in their training data, and our work should not be interpreted as providing perfect or unbiased representations of populations.

Reproducibility Statement

We have taken several steps to facilitate the reproducibility of our work. All proposed algorithms are described in the main text and Appendix D, and the anonymized source code is provided in the supplementary material. Detailed dataset descriptions, data processing steps, hyperparameters, configurations, and evaluation procedures are documented in Appendix B. Main results are reported across multiple random seeds with variance estimates. Complete proofs of our theoretical claims are presented in Appendix E.

References

Aher, G. V., Arriaga, R. I., and Kalai, A. T. Using Large Language Models to Simulate Multiple Humans and Replicate Human Subject Studies. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2023.

Bao, K., Zhang, J., Zhang, Y., Huo, X., Chen, C., and Feng, F. Decoding Matters: Addressing Amplification Bias and Homogeneity Issue in Recommendations for Large Language Models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2024.

Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2021.

Benedetto, L., Aradelli, G., Donvito, A., Lucchetti, A., Cappelli, A., and Buttery, P. Using LLMs to Simulate Students' Responses to Exam Questions. In *Findings of the Association for Computational Linguistics: EMNLP*, 2024.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan,

J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2020.

Bui, N., Nguyen, H. T., Kumar, S., Theodore, J., Qiu, W., Nguyen, V. A., and Ying, R. Mixture-of-Personas Language Models for Population Simulation. In *Findings of the Association for Computational Linguistics (ACL)*, pp. 24761–24778, 2025.

Casper, S., Davies, X., Shi, C., Gilbert, T. K., Scheurer, J., Rando, J., Freedman, R., Korbak, T., Lindner, D., Freire, P., Wang, T. T., Marks, S., Ségerie, C., Carroll, M., Peng, A., Christoffersen, P. J. K., Damani, M., Slocum, S., Anwar, U., Siththaranjan, A., Nadeau, M., Michaud, E. J., Pfau, J., Krasheninnikov, D., Chen, X., Langosco, L., Hase, P., Biyik, E., Dragan, A. D., Krueger, D., Sadigh, D., and Hadfield-Menell, D. Open Problems and Fundamental Limitations of Reinforcement Learning from Human Feedback. *Transactions on Machine Learning Research (TMLR)*, 2023.

Cegin, J., Simko, J., and Brusilovsky, P. ChatGPT to Replace Crowdsourcing of Paraphrases for Intent Classification: Higher Diversity and Comparable Model Robustness. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023.

Choi, H. K. and Li, Y. PICLe: Eliciting Diverse Behaviors from Large Language Models with Persona In-Context Learning. In *Forty-first International Conference on Machine Learning (ICML)*, 2024.

Dominguez-Olmedo, R., Hardt, M., and Mendler-Dünner, C. Questioning the Survey Responses of Large Language Models. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems NeurIPS*, 2024.

Fowler, R. J., Paterson, M. S., and Tanimoto, S. L. Optimal Packing and Covering in the Plane are NP-complete. *Information Processing Letters*, 12(3):133–137, 1981.

Goel, S., Strüber, J., Auzina, I. A., Chandra, K. K., Kumaraguru, P., Kiela, D., Prabhu, A., Bethge, M., and Geiping, J. Great Models Think Alike and this Undermines AI Oversight. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2025.

Ji, B., Duan, X., Qiu, Z., Zhang, T., Li, J., Yang, H., and Zhang, M. Submodular-based In-context Example Selection for LLMs-based Machine Translation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC/COLING)*, pp. 15398–15409, 2024.

Jiang, L., Chai, Y., Li, M., Liu, M., Fok, R., Dziri, N., Tsvetkov, Y., Sap, M., Albalak, A., and Choi, Y. Artificial Hivemind: The Open-Ended Homogeneity of Language Models (and Beyond). In *Proceedings of the Annual Conference on Neural Information Processing Systems (NeurIPS), Datasets and Benchmark Track*, 2025.

Kamath, A. et al. Gemma 3 Technical Report. *CoRR*, abs/2503.19786, 2025.

Kirk, H. R., Vidgen, B., Röttger, P., and Hale, S. A. Personalisation Within Bounds: A Risk Taxonomy and Policy Framework for the Alignment of Large Language Models with Personalised Feedback. *CoRR*, abs/2303.05453, 2023.

Kirk, R., Mediratta, I., Nalmpantis, C., Luketina, J., Hambro, E., Grefenstette, E., and Raileanu, R. Understanding the Effects of RLHF on LLM Generalisation and Diversity. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024.

Kovac, G., Sawayama, M., Portelas, R., Colas, C., Dominey, P. F., and Oudeyer, P. Large Language Models as Superpositions of Cultural Perspectives. *CoRR*, abs/2307.07870, 2023.

Krause, A. and Golovin, D. Submodular Function Maximization. In *Tractability: Practical Approaches to Hard Problems*, pp. 71–104. 2014.

Lahoti, P., Blumm, N., Ma, X., Kotikalapudi, R., Potluri, S., Tan, Q., Srinivasan, H., Packer, B., Beirami, A., Beutel, A., and Chen, J. Improving Diversity of Demographic Representation in Large Language Models via Collective-Critiques and Self-Voting. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023.

Lee, M. H. J., Montgomery, J. M., and Lai, C. K. Large Language Models Portray Socially Subordinate Groups as More Homogeneous, Consistent with a Bias Observed in Humans. In *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2024.

Liu, R., Zhang, G., Feng, X., and Vosoughi, S. Aligning Generative Language Models with Human Values. In *Findings of the Association for Computational Linguistics: NAACL*, 2022.

Liu, R., Sumers, T. R., Dasgupta, I., and Griffiths, T. L. How do Large Language Models Navigate Conflicts between Honesty and Helpfulness? In *Forty-first International Conference on Machine Learning (ICML)*, 2024.

Markel, J. M., Opferman, S. G., Landay, J. A., and Piech, C. GPTeach: Interactive TA Training with GPT-based Students. In *Proceedings of the Tenth ACM Conference on Learning @ Scale (L@S)*, 2023.

McCrae, R. R. and John, O. P. An Introduction to the Five-factor Model and its Applications. *Journal of personality*, 60 2:175–215, 1992.

McInnes, L. and Healy, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *CoRR*, abs/1802.03426, 2018.

Mirzasoleiman, B., Badanidiyuru, A., Karbasi, A., Vondrák, J., and Krause, A. Lazier Than Lazy Greedy. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence (AAAI)*, 2015.

Mohamed, Y., Khan, F. F., Haydarov, K., and Elhoseiny, M. It is Okay to Not Be Okay: Overcoming Emotional Bias in Affective Image Captioning by Contrastive Data Collection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Mohammad, S. M. and Kiritchenko, S. WikiArt Emotions: An Annotated Dataset of Emotions Evoked by Art. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC)*, 2018.

Moskovskiy, D., Pletenev, S., and Panchenko, A. LLMs to Replace Crowdsourcing For Parallel Data Creation? The Case of Text Detoxification. In *Findings of the Association for Computational Linguistics: EMNLP*, 2024.

Nemhauser, G. L., Wolsey, L. A., and Fisher, M. L. An Analysis of Approximations for Maximizing Submodular Set Functions - I. *Math. Program.*, 14(1), 1978.

Nguyen, M. H., Tschiatschek, S., and Singla, A. Large Language Models for In-Context Student Modeling: Synthesizing Student's Behavior in Visual Programming. In *Proceedings of the 17th International Conference on Educational Data Mining (EDM)*, 2024.

Nguyen, M. H., Padurean, V., Gotovos, A., Tschiatschek, S., and Singla, A. Synthesizing High-Quality Programming Tasks with LLM-Based Expert and Student Agents. In *Proceedings of the International Conference on Artificial Intelligence in Education (AIED)*, 2025.

Nguyen, M. H., Tschiatschek, S., and Singla, A. Enhancing Diversity of LLM-Generated Educational Tasks. In *Proceedings of the International Conference on Educational Data Mining (EDM)*, 2026.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P. F., Leike, J., and Lowe, R. Training Language Models to Follow Instructions with Human Feedback. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

Panteli, A. and Boutsinas, B. Improvement of similarity-diversity trade-off in recommender systems based on a facility location model. *Neural Comput. Appl.*, 35(1): 177–189, 2023.

Qian, J., Sun, M., Zhou, S., Zhao, Z., Hun, R., and Chiang, P. Sub-SA: Strengthen In-Context Learning via Submodular Selective Annotation. In *Proceedings of the European Conference on Artificial Intelligence (ECAI)*, volume 392 of *Frontiers in Artificial Intelligence and Applications*, pp. 2034–2041. IOS Press, 2024.

Ramezani, A. and Xu, Y. Knowledge of Cultural Moral Norms in Large Language Models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*, 2023.

Rammstedt, B. and John, O. P. Measuring Personality in One Minute or Less: A 10-item Short Version of the Big Five Inventory in English and German. *Journal of Research in Personality*, 41(1):203–212, 2007.

Salewski, L., Alaniz, S., Rio-Torto, I., Schulz, E., and Akata, Z. In-Context Impersonation Reveals Large Language Models' Strengths and Biases. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2023.

Santurkar, S., Durmus, E., Ladhak, F., Lee, C., Liang, P., and Hashimoto, T. Whose Opinions Do Language Models Reflect? In *International Conference on Machine Learning (ICML)*, 2023.

Scholz, N., Nguyen, M. H., Singla, A., and Nagashima, T. Partnering with AI: A Pedagogical Feedback System for LLM Integration Into Programming Education. In *ECTEL*, 2025.

Shumailov, I., Shumaylov, Z., Zhao, Y., Papernot, N., Anderson, R. J., and Gal, Y. AI Models Collapse When Trained on Recursively Generated Data. *Nature*, 631 (8022):755–759, 2024.

Singla, A., Horvitz, E., Kamar, E., and White, R. Stochastic Privacy. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence (AAAI)*, 2014.

Sorensen, T., Moore, J., Fisher, J., Gordon, M. L., 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 *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*, 2024.

Sun, S., Lee, E., Nan, D., Zhao, X., Lee, W., Jansen, B. J., and Kim, J. Random Silicon Sampling: Simulating Human Sub-Population Opinion Using a Large Language Model Based on Group-Level Demographic Information. *CoRR*, abs/2402.18144, 2024.

Sun, X., Li, X., Li, J., Wu, F., Guo, S., Zhang, T., and Wang, G. Text Classification via Large Language Models. In *Findings of the Association for Computational Linguistics: EMNLP*, 2023.

Tan, W. R., Chan, C. S., Aguirre, H., and Tanaka, K. Improved ArtGAN for Conditional Synthesis of Natural Image and Artwork. *IEEE Transactions on Image Processing*, 28(1):394–409, 2019.

Tao, Y., Viberg, O., Baker, R. S., and Kizilcec, R. F. Cultural Bias and Cultural Alignment of Large Language Models. *PNAS Nexus*, 3(9):pgae346, 2024.

Trust, P. and Minghim, R. Query-Focused Submodular Demonstration Selection for In-Context Learning in Large Language Models. In *Proceedings of the Irish Conference on Artificial Intelligence and Cognitive Science (AICS)*, pp. 1–8. IEEE, 2023.

Verter, V. Uncapacitated and Capacitated Facility Location Problems. In *Foundations of Location Analysis*, pp. 25–37. New York, NY, 2011.

Wang, Z., Lamb, A., Saveliev, E., Cameron, P., Zaykov, Y., Hernández-Lobato, J. M., Turner, R. E., Baraniuk, R. G., Barton, C., Jones, S. P., Woodhead, S., and Zhang, C. Diagnostic Questions: The NeurIPS 2020 Education Challenge. *CoRR*, abs/2007.12061, 2020.

Wang, Z., Zhang, K., Liu, Z., Qiao, R., Kang, X., Jiang, J., Chuang, Y., Wang, W., Zettlemoyer, L., Nie, J., and Ji, H. OASIS: Open Agent Social Interaction Simulations with One Million Agents. *CoRR*, abs/2411.11581, 2024.

Wenger, E. and Kenett, Y. N. We're Different, We're the Same: Creative Homogeneity Across LLMs. *CoRR*, abs/2501.19361, 2025.

Wu, F., Black, E., and Chandrasekaran, V. Generative Monoculture in Large Language Models. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2025.

Wu, Q., Bansal, G., Zhang, J., Wu, Y., Li, B., Zhu, E., Jiang, L., Zhang, X., Zhang, S., Liu, J., Awadallah, A. H., Whitaker, R., Chen, R., Hoover, B., Hebert, M., Burger, D., Galley, M., Horvitz, E., Beirami, A., and Wang, H. AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation. In *Proceedings of the First Conference on Language Modeling (COLM)*, 2024.

Xie, C., Chen, C., Jia, F., Ye, Z., Lai, S., Shu, K., Gu, J., Bibi, A., Hu, Z., Jurgens, D., Evans, J., Torr, P., Ghanem, B., and Li, G. Can Large Language Model Agents Simulate Human Trust Behavior? In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2024.

Yoon, S., He, Z., Echterhoff, J. M., and McAuley, J. J. Evaluating Large Language Models as Generative User Simulators for Conversational Recommendation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), NAACL*, 2024.

Yu, Y., Zhuang, Y., Zhang, J., Meng, Y., Ratner, A. J., Krishna, R., Shen, J., and Zhang, C. Large Language Model as Attributed Training Data Generator: A Tale of Diversity and Bias. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2023.

Zhang, R., Li, Y., Ma, Y., Zhou, M., and Zou, L. LLMaAA: Making Large Language Models as Active Annotators. In *Findings of the Association for Computational Linguistics: EMNLP*, 2023.

Zhang, Z., Zhang-Li, D., Yu, J., Gong, L., Zhou, J., Hao, Z., Jiang, J., Cao, J., Liu, H., Liu, Z., Hou, L., and Li, J. Simulating Classroom Education with LLM-Empowered Agents. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, 2025.

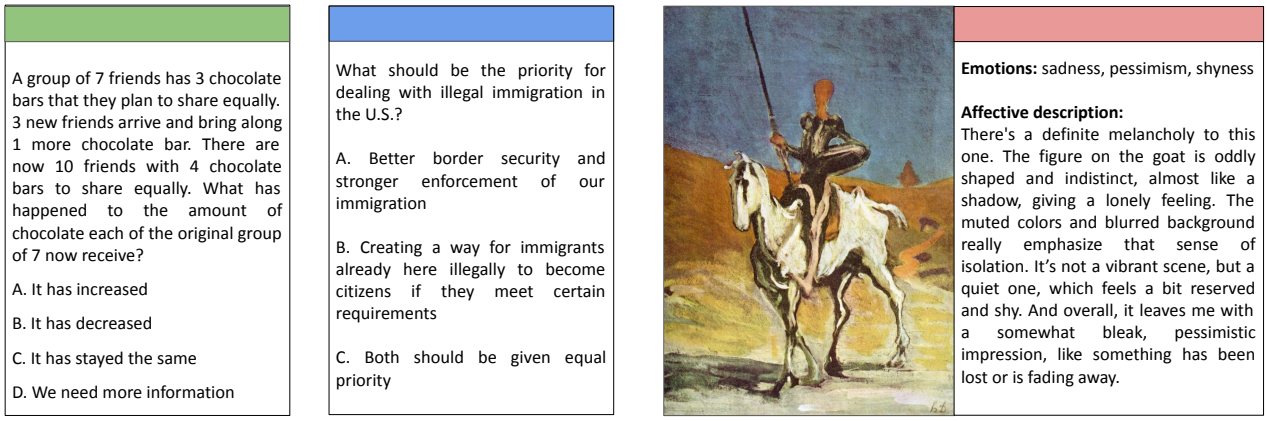
Appendix

A. Table of Contents

In this section, we briefly describe the content provided in the paper's appendices.

- Section B provides more details about the experimental setup, including datasets and prompts used for each domain, computational resources, and method configurations.
- Section C provides results on multiple runs with different context sizes K and random seeds.
- Section D provides pseudocode for the proposed algorithms.
- Section E provides detailed proofs of our theorems and propositions.

B. Additional Experimental Setup



(a) Task in EEDI *(b)* Task in OpinionQA *(c)* Painting in WikiArt with an annotation example.

Figure 5. Examples of tasks in our experiments.

B.1. EEDI

Dataset. We use the EEDI math dataset (Wang et al., 2020), which provides data on students' answers to mathematics questions on the Eedi platform (see Figure 5a for an example). Representing students with varying levels of skills and misconceptions can benefit both teachers and learners, allowing for virtual pre-testing and instructional practice in a safe environment. The questions are multiple-choice with four answer choices presented as images, and we select 40 that can be converted to text and have a large number of student responses, chosen to maximize both the number of questions answered per student and the number of students answering each question. The selected questions cover four concepts: Fractions, Negative Numbers, Mental Multiplication and Division, and Simplifying Expressions. Half of the questions in each concept are used for training and the other half for testing.

Prompt for agent. Each agent is given a set of K demonstrations, which are pairs of math questions and example answers provided by the real students (cf. Figure 6). Then, we ask the agent to analyze the given answers and predict how the student would answer a new question.

Figure 6: [EEDI] Prompt for Agent with K demonstrations

[User message]
{question_1}
{example_answer_of_question_1}
.
.
.
{question_K}
{example_answer_of_question_K}
Evaluate whether the student's previous answers reveal any misconceptions. If so, analyze those misconceptions before proceeding. If not, directly predict how the student would answer the following question.
{question}

Embeddings. Each student or agent is represented by a binary embedding vector indicating their correct and incorrect answers to the math questions. For example, $[1, 0, \ldots, 1]$ represents [correct, incorrect, ..., correct]. This vector summarizes the behavior of the student/agent on a set of math questions. Distances between embeddings are computed using L1 distance, which in this case is equivalent to the Hamming distance. Intuitively, it counts the number of questions on which two students' answers differ, reflecting their performance difference.

B.2. OpinionQA

Dataset. We use questions and answers from the American Trends Panel W92 survey data, which was used in (Santurkar et al., 2023) (see Figure 5b for an example). In this application, LLMs function as surrogates for crowdworkers, providing responses that reflect the diverse opinions and beliefs of different population segments. This survey includes 77 multiple-choice questions related to politics and responses from over $10,000$ respondents across the US. The answer choices typically have an ordinal structure (e.g., ranging from "A great deal" to "Not at all") and we reuse the mapping from answer choices to ordinal values from (Santurkar et al., 2023). In addition, the survey data includes demographic information of the people, including ideology, political party and region. We sample $N = 500$ people and take their answers to create our dataset. We use the demographic information of these people for analysis purposes only.

Prompt for agent. Each agent is given a set of K demonstrations, which are pairs of questions and example answers provided by the real survey respondents (cf. Figure 7). Then, we ask the agent to act as the human who has given the answers above and to answer a new question.

Figure 7: [OpinionQA] Prompt for Agent with K demonstrations

[User message]
{question_1}
{example_answer_of_question_1}
.
.
.
{question_K}
{example_answer_of_question_K}
Act as the human who has given the answers above. Answer the following question.
{question}

Response generation. Unlike some prior work that extracts next-token probabilities for answer choices (Dominguez-Olmedo et al., 2024), we generate full text responses and then parse the output to identify the selected answer. This allows the model to attend to the semantic content of the answer choices rather than relying solely on positional token probabilities. However, we acknowledge that this approach may still be subject to ordering biases, which we leave for future work to investigate.

Embeddings. For each question, answer choices are mapped to ordinal values and normalized to the range $[-1, 1]$. Each human or agent is then represented by a vector embedding of their responses. For example, $[-1, 0.5, \ldots, 1]$ corresponds to answer choices such as ['worse', 'Increased somewhat', ..., 'Greatly expand on current government services']. This vector summarizes the behavior of the person/agent on a set of survey questions. Distances between embeddings are computed using L2 distance. This distance captures the extent of differences in opinions and beliefs expressed in their answers.

B.3. Wikiart

Figure 8: [Wikiart] Prompt for Synthetic Human

[System message] Act as a human who has taken the following personality test. Be mindful of how you focus your attention and the way you express yourself through language.
I see myself as someone who is generally trusting: {answer}
I see myself as someone who tends to be lazy: {answer}
I see myself as someone who is relaxed, handles stress well: {answer}
I see myself as someone who has few artistic interests: {answer}
I see myself as someone who is outgoing, sociable: {answer}
I see myself as someone who tends to find fault with others: {answer}
I see myself as someone who does a thorough job: {answer}
I see myself as someone who gets nervous easily: {answer}
I see myself as someone who has an active imagination: {answer}
[User message] What emotions does the following painting evoke? Choose from the following list of emotions: gratitude, happiness, humility, love, optimism, trust, anger, arrogance, disgust, fear, pessimism, regret, sadness, shame, agreeableness, anticipation, disagreeableness, shyness, surprise. Provide a short explanation referencing specific details from the painting. Respond in JSON format with two keys: "emotions" and "explanation".
{painting}

Figure 9: [Wikiart] Prompt for extracting embeddings for an annotator

[User message]
Response_1: annotation_of_painting_1
.
.
.
Response_q: annotation_of_painting_q

Figure 10: [Wikiart] Prompt for Agent with K demonstrations

[User message]
{painting_1}
{annotation_of_painting_1}
.
.
.
{painting_K}
{annotation_of_painting_K}
Act as the human who has given the answers above. What emotions does the following painting evoke? Choose from the following list of emotions: gratitude, happiness, humility, love, optimism, trust, anger, arrogance, disgust, fear, pessimism, regret, sadness, shame, agreeableness, anticipation, disagreeableness, shyness, surprise. Provide a short explanation referencing specific details from the painting. Respond in JSON format with two keys: "emotions" and "explanation".
{painting}

Dataset. Figure 5c shows an example painting with an annotation. LLMs can be used here as surrogates for crowdworkers in annotation tasks where diverse perspectives and nuances in language use are encouraged. We use LLMs conditioned on the Big Five personality traits (McCrae & John, 1992), namely Extraversion, Agreeableness, Conscientiousness, Emotional Stability, and Openness, to act as annotators. Each annotator is specified by responses to a 10-item BFI questionnaire (Rammstedt & John, 2007), which we place in the system message to instruct the LLM to act as a human who has taken the test (cf. Figure 8). Responses take one of five values (Disagree Strongly, Disagree a Little, Neither Agree nor Disagree,

Agree a Little, Agree Strongly), which we convert into trait scores using the formula in (Rammstedt & John, 2007) and rescale to $[0, 5]$, where higher values indicate stronger expression of the trait. To create a diverse set of annotators, we sample responses from normal distributions of trait values with varying means and standard deviations. These system messages condition the LLM to act with distinct personalities, and we use Gemma3-27B (Kamath et al., 2025) to generate their answers.

Embeddings. To create an embedding for each human or agent, we concatenate their answers on the train/test tasks into a single prompt (cf. Figure 9), pass it through a language model (Gemma3-12B (Kamath et al., 2025)), and extract the last hidden state with mean pooling to obtain the embedding. We then reduce the dimensionality to 64 using PCA. The resulting continuous embedding vector (e.g., $[0.1032, -0.2211, \ldots, -0.7715]$) summarizes the behavior of the human annotator/agent on a set of tasks. Distances between these embeddings are computed using L2 distance, which captures differences in annotators' perspectives and language behaviors across tasks.

Prompt for agent. Each agent is given a set of K demonstrations, which are pairs of paintings and example annotations provided by the annotators (cf. Figure 10). Then, we ask the agent to act as the annotator who has given the answers above and to annotate a new painting. The LLM agent is given a list of emotions (from (Mohammad & Kiritchenko, 2018)) to choose from, and it must provide a short explanation referencing specific details from the painting.

B.4. Baselines Details

We provide additional details on the implementation of the three baselines (ATTRPROMPT, PICLE, MOP) adapted to our setting.

AttrPrompt (Yu et al., 2023) generates diverse synthetic data through a human-AI collaboration process. First, we prompt an LLM to identify key attribute dimensions (e.g., skill level, political affiliation) that influence behavioral diversity in the domain and suggest possible values for each. We then create M agent profiles by randomly sampling values for each assigned attribute dimension. Finally, for each agent profile, we generate K synthetic demonstrations by prompting the LLM to answer questions as if it were a person with the sampled profile. The set of K synthetic demonstrations is used in the agent's prompt.

PICLe (Choi & Li, 2024) selects effective personas from human demonstrations using a likelihood-ratio scoring mechanism. For each agent, we sample a human $h \in \mathcal{H}$ and fine-tune a base LLM on all their demonstrations using LoRA (rank $r = 8$, alpha $\alpha = 32$) for 4 epochs with AdamW (learning rate 2e-5). We then rank the demonstrations by the difference $\Delta = \log P_{\text{FT}}(a_i|q_i) - \log P_{\text{Base}}(a_i|q_i)$, where P_{FT} and P_{Base} denote the likelihood under the fine-tuned and base models, respectively. The top-K demonstrations with the highest Δ scores are selected and used in the agent's prompt.

Mixture-of-Personas (MoP) (Bui et al., 2025) aligns LLM responses with target population distributions using a hierarchical mixture model. We adapt this method to construct agents by first clustering all human demonstrations in the training set into M clusters using K-means on their embeddings. For each cluster, we randomly sample demonstrations and prompt an LLM to generate a concise persona description. We then train a gating network to estimate the relevance of each demonstration relative to a given context and the generated persona. To construct the final agent for the cluster, we compute the average embedding of the contexts in the cluster and select the K demonstrations that maximize the gating score relative to this average context.

B.5. Additional Details

Stochastic Greedy. Accelerated evaluation in $\text{REPPOP}_{\text{demo}}$ is achieved via a stochastic greedy variant that samples a subset of α demonstration candidates from the pool $\mathcal{D}_{\mathcal{H}}^{\mathcal{T}}$. We set $\alpha = 100$ for WikiArt and EEDI, and $\alpha = 1000$ for OpinionQA, chosen based on the scale of $\mathcal{D}_{\mathcal{H}}^{\mathcal{T}}$.

Generation Parameters. All agents use a decoding temperature of 1.0 to allow sufficient stochasticity for expressing diverse behaviors reflective of the population.

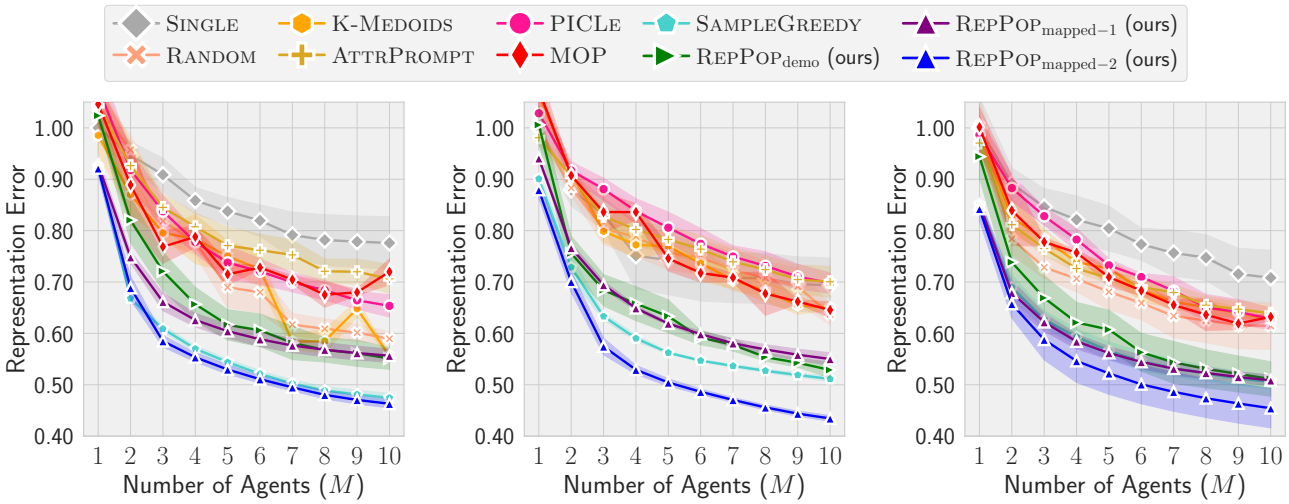
B.6. Resources

We use machines with 2 x Intel Xeon Gold 5317 for all experiments. We use 1 x NVIDIA H100 80GB for experiments on EEDI and OpinionQA datasets, and 1 x NVIDIA H200 141GB for Wikiart dataset. We use 8xH200 141GB with parallelization for our runtime analysis.

C. Additional Experimental Results

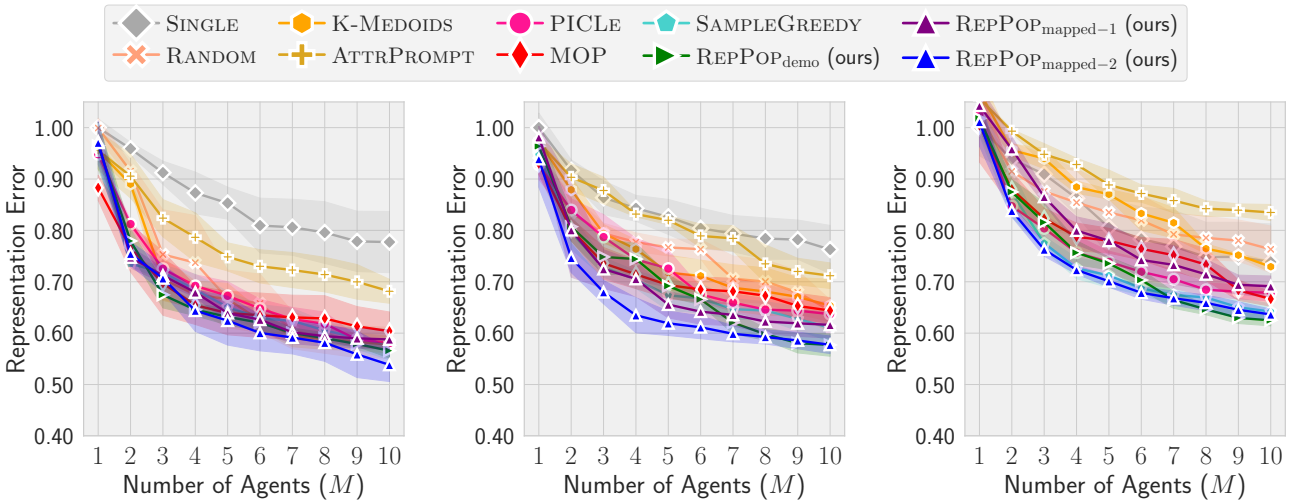
C.1. Representation Error when Varying K

In this section, we show results on different context sizes $K = 1, 3, 5$ for each dataset. We report the mean and standard error (shown as error bars) computed over three seeds. Overall, we observe similar trends where our methods outperform other baselines across all datasets.



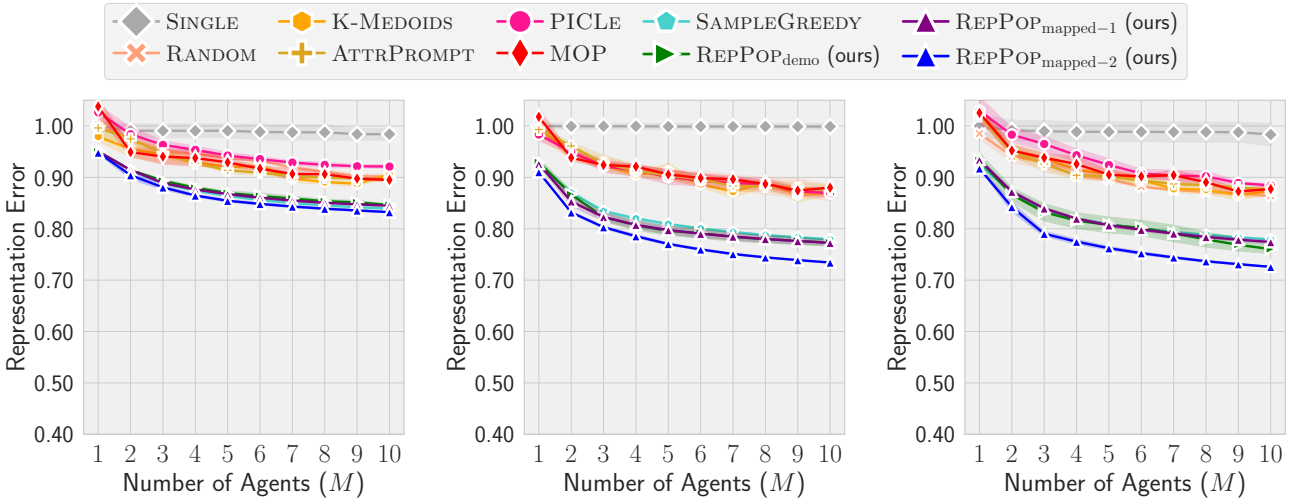
(a) $K = 1$ *(b)* $K = 3$ *(c)* $K = 5$

Figure 11. [EEDI-Train] Representation error.



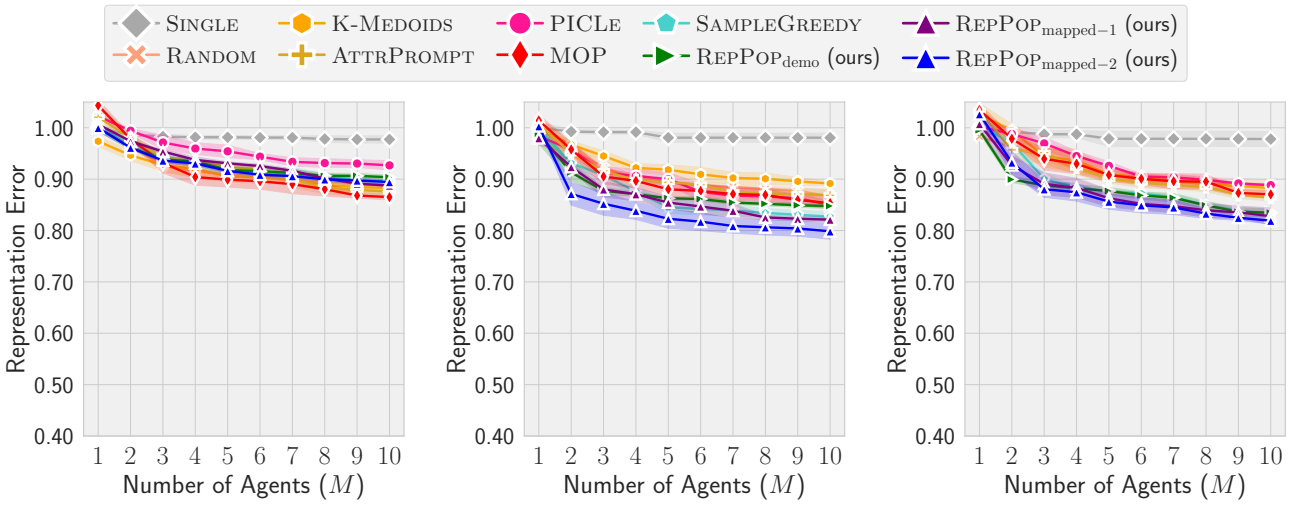
(a) $K = 1$ *(b)* $K = 3$ *(c)* $K = 5$

Figure 12. [EEDI-Validation] Representation error.



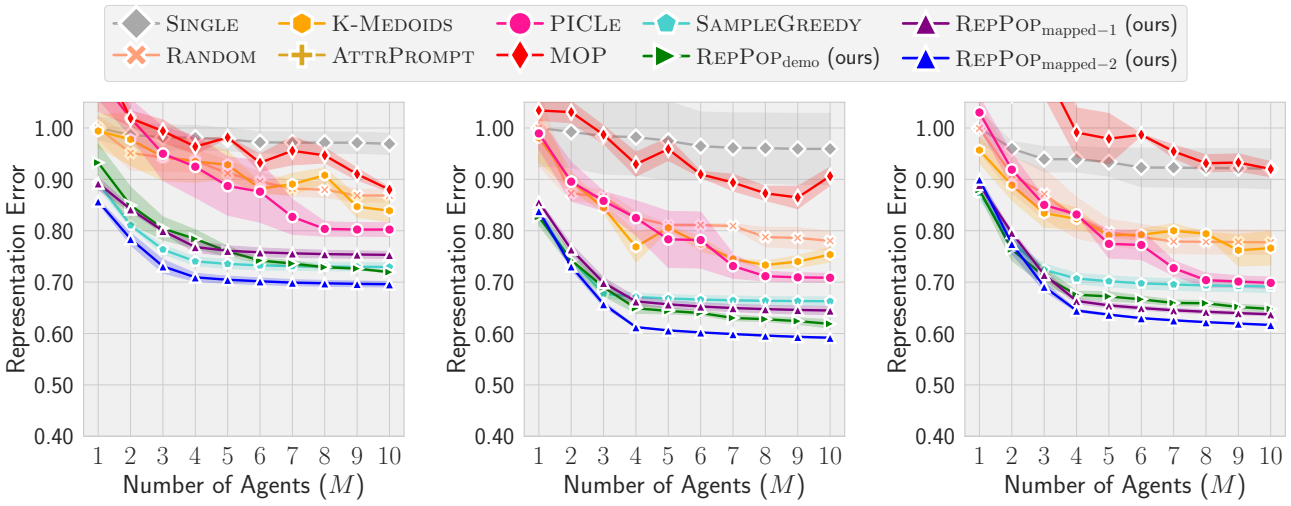
(a) $K = 1$ *(b)* $K = 3$ *(c)* $K = 5$

Figure 13. [OpinionQA-Train] Representation error.



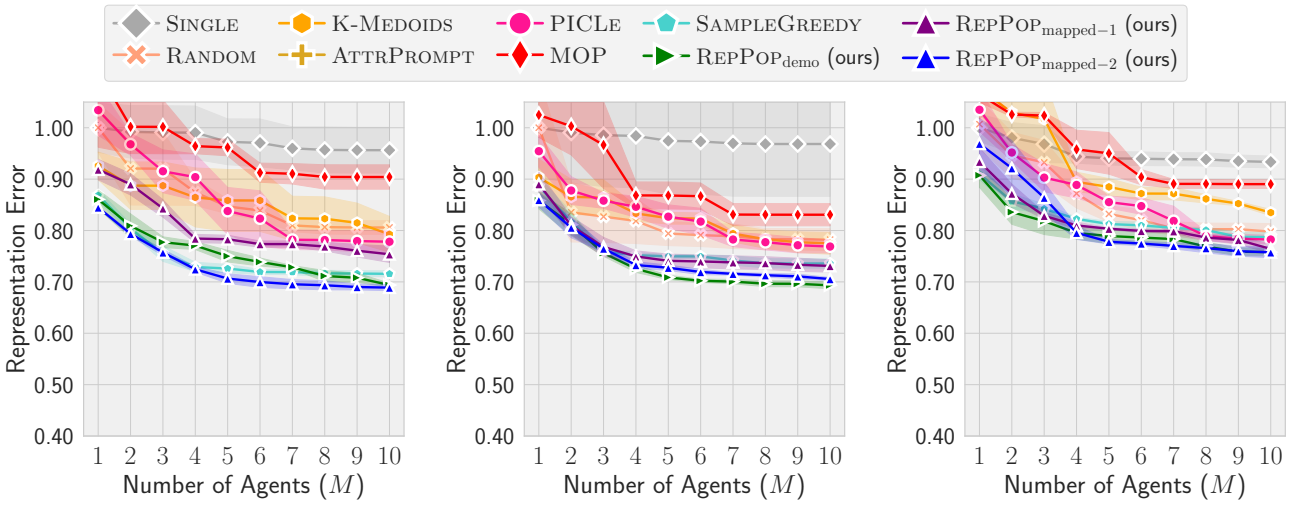
(a) $K = 1$ *(b)* $K = 3$ *(c)* $K = 5$

Figure 14. [OpinionQA-Validation] Representation error.



(a) $K = 1$ *(b)* $K = 3$ *(c)* $K = 5$

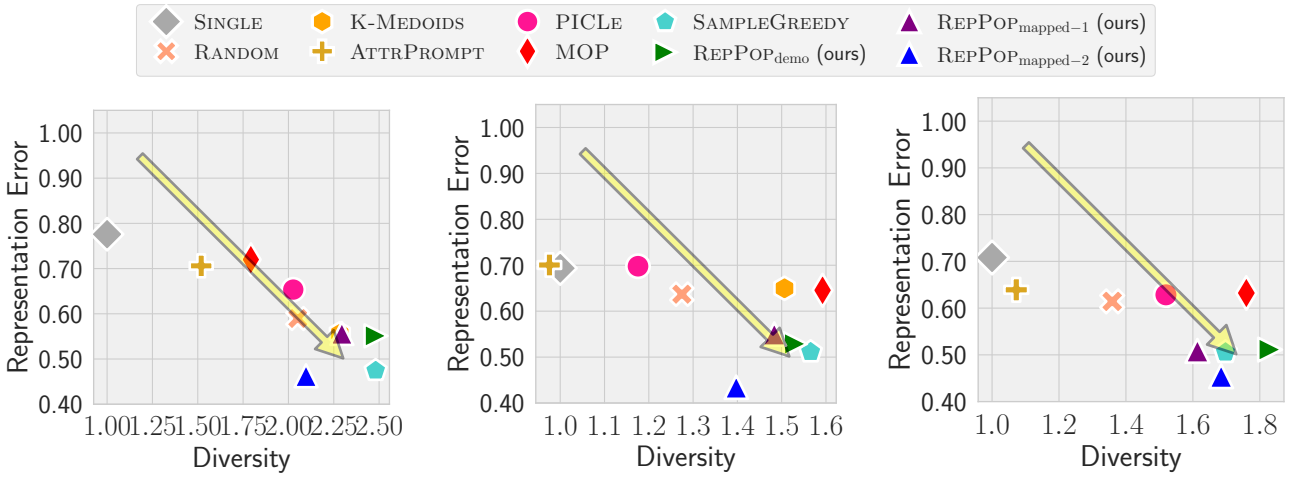
Figure 15. [Wikiart-Train] Representation error.



(a) $K = 1$ *(b)* $K = 3$ *(c)* $K = 5$

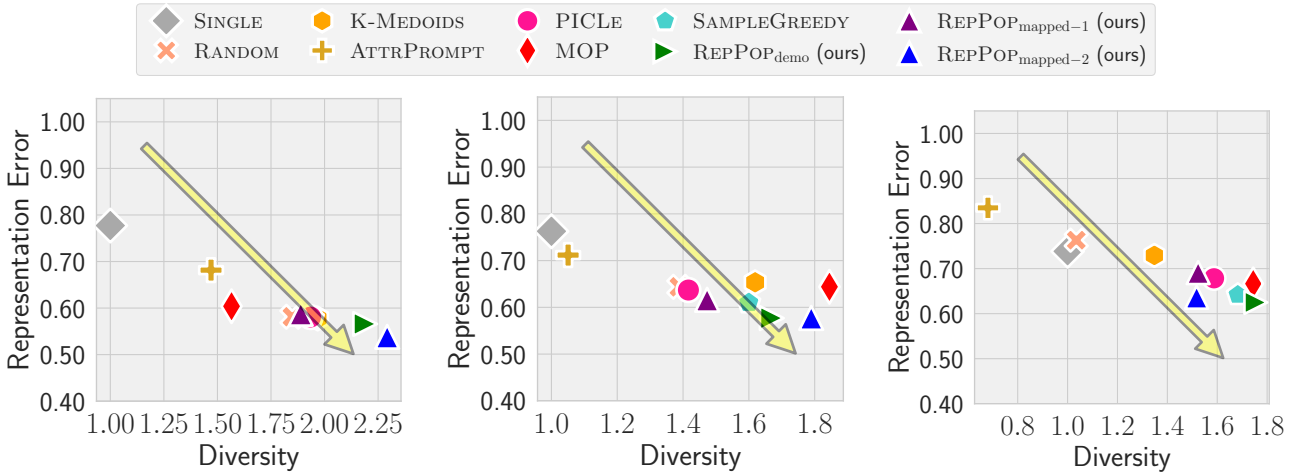
Figure 16. [Wikiart-Validation] Representation error.

C.2. Diversity and Representation Error when Varying K



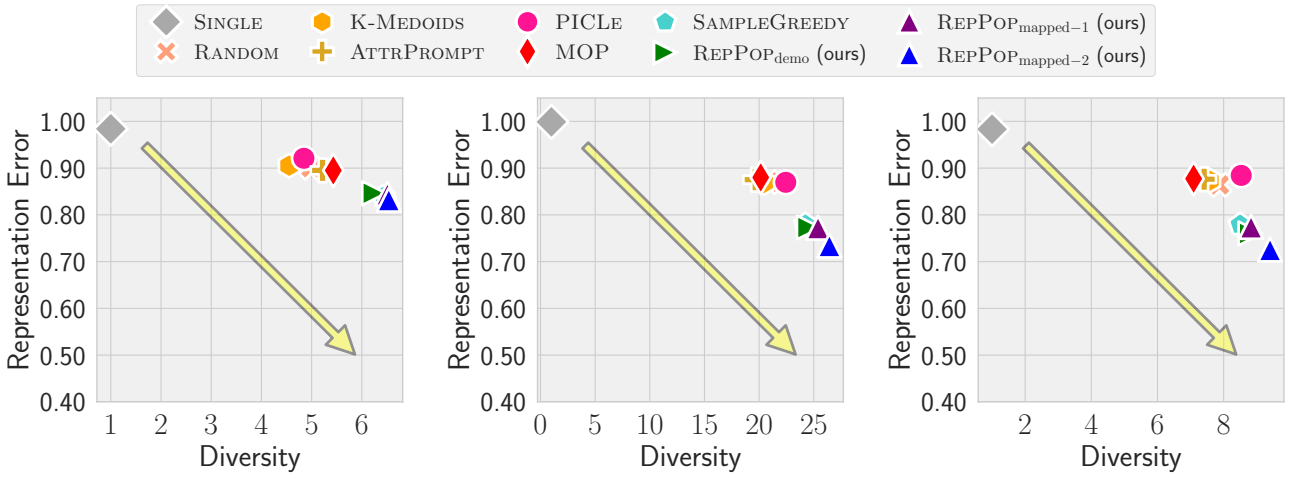
(a) $K = 1$ *(b)* $K = 3$ *(c)* $K = 5$

Figure 17. [EEDI-Train] Diversity and representation error.



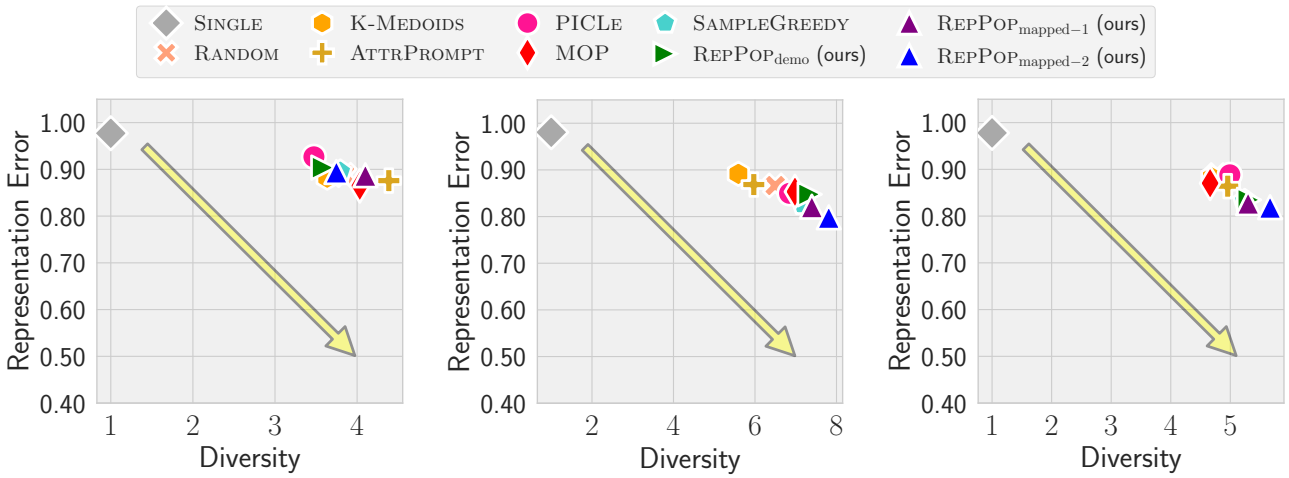
(a) $K = 1$ *(b)* $K = 3$ *(c)* $K = 5$

Figure 18. [EEDI-Validation] Diversity and representation error.



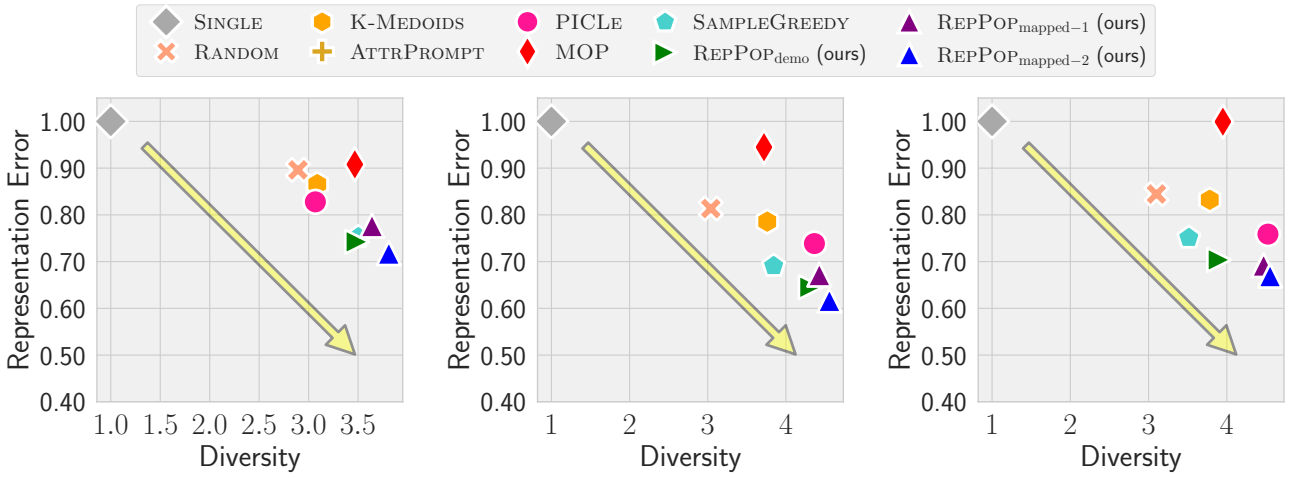
(a) $K = 1$ *(b)* $K = 3$ *(c)* $K = 5$

Figure 19. [OpinionQA-Train] Diversity and representation error.



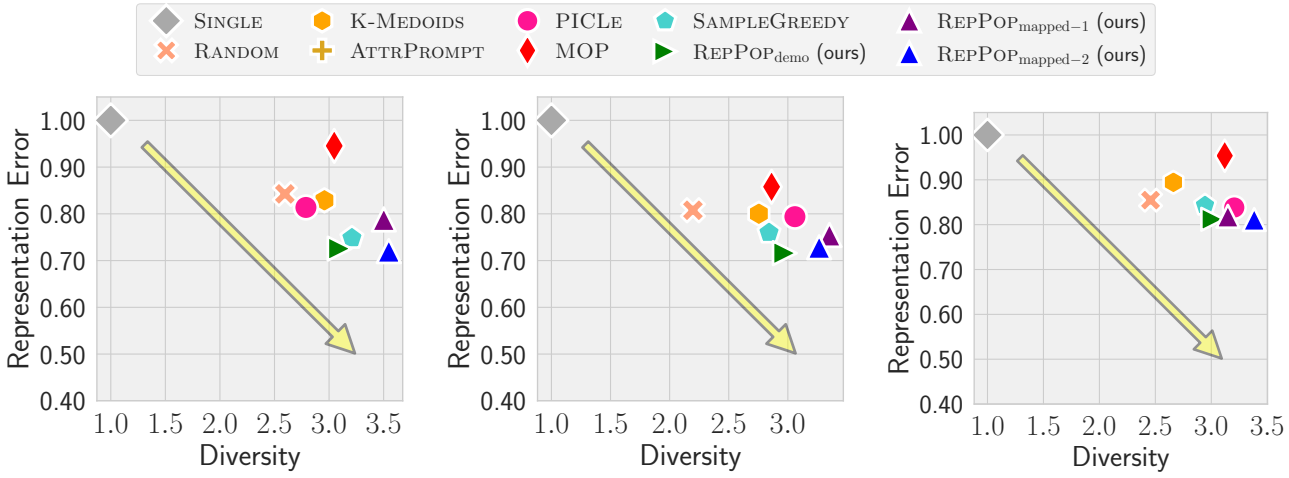
(a) $K = 1$ *(b)* $K = 3$ *(c)* $K = 5$

Figure 20. [OpinionQA-Validation] Diversity and representation error.



(a) $K = 1$ *(b)* $K = 3$ *(c)* $K = 5$

Figure 21. [Wikiart-Train] Diversity and representation error.



(a) $K = 1$ *(b)* $K = 3$ *(c)* $K = 5$

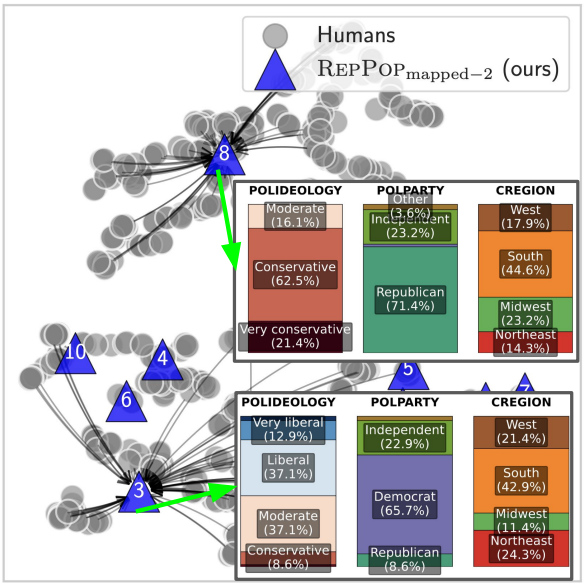
Figure 22. [Wikiart-Validation] Diversity and representation error.

C.3. Agent Behavior Analysis

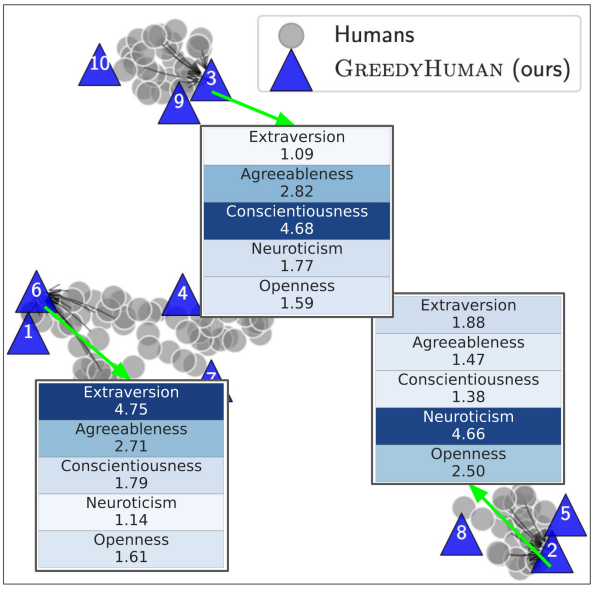
OpinionQA. We select two agents and show the aggregated self-declared metadata of the humans they represent in Figure 23a. For each task in the validation set, we map answer choices to political ideologies and party affiliations. For each agent, we compute the distribution over these labels across all questions. For example, for Agent 3 we obtain the following distributions: political ideologies — Very liberal (13.3%), Liberal (40%), Moderate (13.3%), Conservative (26.7%), Very conservative (6.2%); political parties — Independent (13.3%), Democrat (53.3%), Republican (33.3%). These distributions mirror the actual human demographic patterns for Agent 3 shown in Figure 23a, despite the fact that such metadata was not used in constructing the agents. These findings highlight their ability to capture and reproduce meaningful population-level diversity.

WikiArts. We select three agents and visualize the annotators they represent along with their aggregated traits (cf. Figure 23b). For example, Agent 2 represents annotators with high neuroticism and low conscientiousness. To validate whether the agents exhibit behaviors consistent with the annotators they represent, we ask each constructed agent to complete the 10-item BFI test (Rammstedt & John, 2007), and their responses reveal traits that align with the average traits of the corresponding humans. For instance, Agent 2 in Figure 23b obtains the following trait scores: Extraversion (1.0), Agreeableness (1.5),

Conscientiousness (1.38), Neuroticism (4.66), and Openness (2.5), aligned with the annotators it represents in Figure 23b.



(a) OpinionQA. Each agent represents a group of humans with particular distributions of political ideologies, parties, and regions.



(b) WikiArts. Each agent represents a group of annotators with particular personality traits.

Figure 23. 2D embeddings of humans and agents constructed by REPPOP$_{\text{mapped-2}}$ on tasks in $\mathcal{T}_{\text{train}}$ using UMAP (McInnes & Healy, 2018). We provide examples of aggregated metadata (in the boxes) of humans represented by agents (connections are denoted by black arrows). They are not used for constructing agents and used only for analysis. Our method REPPOP$_{\text{mapped-2}}$ constructs agents to cover different human behaviors, collectively representing the human population.

C.4. Varying Generative and Embedding Models

We investigate how different methods perform when using different underlying generative models for the agents (results in Table 3). We evaluate our methods against several baselines, including the best baseline SAMPLEGREEDY.

Method	Model					
	Phi-4-mini (4B)	**Phi-4 (14B)**	**Qwen-3 (14B)**	**Gemma-3 (27B)**	**Qwen-3 (32B)**	**Llama-3.1 (70B)**
SINGLE	0.81	0.69	0.92	0.76	0.79	0.84
RANDOM	0.82	0.63	0.78	0.75	0.73	0.84
K-MEDOIDS	0.86	0.63	0.82	0.77	0.77	0.69
SAMPLEGREEDY	0.81	0.64	0.74	0.73	0.72	0.67
REPPOP$_{\text{demo}}$	0.77	**0.53**	0.72	**0.67**	**0.69**	**0.64**
REPPOP$_{\text{mapped-1}}$	0.79	0.62	0.72	0.69	0.73	0.69
REPPOP$_{\text{mapped-2}}$	**0.74**	0.58	**0.66**	0.69	0.74	0.69

Table 3. Performance when using other generative models. Representation error on the EEDI dataset with $M = 10$ and $K = 3$ across model families and varied sizes (4B–70B). Results are reported on the validation set, with bold numbers indicating the lowest error. Our proposed methods consistently outperform baselines, demonstrating lower representation error across all tested models and highlighting the robustness of our framework independent of the underlying model choice.

We evaluate our methods on the WikiArt dataset using a smaller and more efficient bidirectional encoder (gte-base-en-v1.5). The results in Table 4 confirm that our methods remain effective even when relying on a much smaller embedding model.

Method	gte-base-en-v1.5 (137M)
SINGLE	0.87
RANDOM	0.85
K-MEDOIDS	0.84
SAMPLEGREEDY	0.83
REPPOP$_{\text{demo}}$ (ours)	**0.76**
REPPOP$_{\text{mapped-1}}$ (ours)	0.81
REPPOP$_{\text{mapped-2}}$ (ours)	0.83

Table 4. Comparison of methods on the WikiArt dataset using the gte-base-en-v1.5 embedding model. We report the representation error on the test set with context size $K = 3$ and agent set size $M = 10$.

D. Algorithms

D.1. Pseudocode of REPPOP$_{\text{demo}}$

Algorithm 1 REPPOP$_{\text{demo}}$

Input: Human set $\mathcal{H}$, human demonstrations $\mathcal{D}_{\mathcal{H}}^{\mathcal{T}}$, number of agents to select M, context size K
Output: A set of representative agents $L \subseteq \mathcal{L}$ with $|L| \leq M$
Initialize $L \leftarrow \emptyset$
for $i = 1$ to M **do**
  Initialize $\Omega \leftarrow \emptyset$
  for $k = 1$ to K **do**
    $demo^* \leftarrow \arg\max_{demo \in \mathcal{D}_{\mathcal{H}}^{\mathcal{T}} \setminus \Omega} f(L \cup \{l_{\Omega \cup \{demo\}}\}) - f(L)$
    $\Omega \leftarrow \Omega \cup \{demo^*\}$
  end for
  $L \leftarrow L \cup \{l_\Omega\}$
end for
return L

D.2. Pseudocode of REPPOP$_{\text{mapped-1}}$ and REPPOP$_{\text{mapped-2}}$

Algorithm 2 REPPOP$_{\text{mapped-1}}$ and REPPOP$_{\text{mapped-2}}$

Input: Human set $\mathcal{H}$, human demonstrations $\mathcal{D}_{\mathcal{H}}^{\mathcal{T}}$, number of agents to select M, context size K
Output: A set of representative agents $L \subseteq \mathcal{L}$ with $|L| = M$
Initialize $\tilde{\mathcal{L}} \leftarrow \emptyset$, $L \leftarrow \emptyset$
for each human $h \in \mathcal{H}$ **do**
  Create agent l_h using K demonstrations from $\mathcal{D}_h^{\mathcal{T}}$
  $\tilde{\mathcal{L}} \leftarrow \tilde{\mathcal{L}} \cup \{l_h\}$
end for
for $i = 1$ to M **do**
  $l^* \leftarrow \arg\max_{l \in \tilde{\mathcal{L}} \setminus L} f(L \cup \{l\}) - f(L)$
  $L \leftarrow L \cup \{l^*\}$
end for
return L

E. Proofs

E.1. Proof of Proposition 4.1

Proposition 4.1 (NP-Hardness). The problem of selecting an optimal subset $L^* \subseteq \mathcal{L}$ of size M that maximizes $f(L)$ is NP-hard.

Proof. We show NP-hardness through a reduction from the k-facility location problem which extends the uncapacitated facility location problem (UFLP) by including a constraint on the maximum number of facilities. The problem is known to be NP-hard. (Fowler et al., 1981)

Consider an instance of the k-facility location problem with a set of potential facility locations $\mathcal{F} = \{f_1, f_2, \ldots, f_{n_F}\}$, a set of customers $\mathcal{C} = \{c_1, c_2, \ldots, c_{n_C}\}$, service costs $d(f, c)$ representing the cost of serving customer c from facility f, and a cardinality constraint M. The objective is to select a subset $F \subseteq \mathcal{F}$ of M facilities to minimize the sum of service costs $\sum_{c \in \mathcal{C}} \min_{f \in F} d(f, c)$.

We construct a corresponding instance of our representative agent selection problem as follows: set $\mathcal{H} = \mathcal{C}$ (each customer corresponds to a human); set $\mathcal{L} = \mathcal{F}$ (each potential facility location corresponds to a potential agent); define $\text{dist}(\mathbf{e}_h, \mathbf{e}_l) = d(l, h)$ for each human $h \in \mathcal{H}$ and agent $l \in \mathcal{L}$; and set M to be the number of facilities we wish to open.

Under this construction, our objective function becomes:

$$\begin{aligned} f(L) &= \frac{1}{|\mathcal{H}|} \sum_{h \in \mathcal{H}} \left[D_{\max} - \min_{l \in L} d(l, h) \right] \\ &= \frac{1}{|\mathcal{H}|} \left(\sum_{h \in \mathcal{H}} D_{\max} \right) - \frac{1}{|\mathcal{H}|} \sum_{h \in \mathcal{H}} \min_{l \in L} d(l, h) \\ &= D_{\max} - \frac{1}{|\mathcal{H}|} \sum_{h \in \mathcal{H}} \min_{l \in L} d(l, h) \end{aligned}$$

Since $D_{\max}$ is a constant and $\frac{1}{|\mathcal{H}|}$ is a positive constant, maximizing $f(L)$ is equivalent to minimizing $\sum_{h \in \mathcal{H}} \min_{l \in L} d(l, h)$, which is the objective of the k-facility location problem.

Therefore, if the representative agent selection problem could be solved in polynomial time, the k-facility location problem could also be solved in polynomial time. Since the k-facility location problem is NP-hard, the representative agent selection problem must also be NP-hard. □

E.2. Proof of Proposition 4.2

Proposition 4.2 (Submodularity of the Objective Function $f(L)$). The objective function

$$f(L) = \frac{1}{|\mathcal{H}|} \sum_{h \in \mathcal{H}} \left[D_{\max} - \min_{l \in L} \text{dist}(\mathbf{e}_h, \mathbf{e}_l) \right]$$

is submodular.

Proof. A set function f is submodular if for all $A \subseteq B \subseteq \mathcal{L}$ and for all $l' \in \mathcal{L} \setminus B$, we have $f(A \cup \{l'\}) - f(A) \geq f(B \cup \{l'\}) - f(B)$.

Let $A \subseteq B \subseteq \mathcal{L}$ and $l' \in \mathcal{L} \setminus B$. The marginal gain of adding l' to A is:

$$f(A \cup \{l'\}) - f(A) = \frac{1}{|\mathcal{H}|} \sum_{h \in \mathcal{H}} \left[D_{\max} - \min_{l \in A \cup \{l'\}} \text{dist}(\mathbf{e}_h, \mathbf{e}_l) - \left(D_{\max} - \min_{l \in A} \text{dist}(\mathbf{e}_h, \mathbf{e}_l) \right) \right]$$

This simplifies to:

$$f(A \cup \{l'\}) - f(A) = \frac{1}{|\mathcal{H}|} \sum_{h \in \mathcal{H}} \left[\min_{l \in A} \text{dist}(\mathbf{e}_h, \mathbf{e}_l) - \min_{l \in A \cup \{l'\}} \text{dist}(\mathbf{e}_h, \mathbf{e}_l) \right]$$

Define $\mathcal{H}_A^{l'} = \{h \in \mathcal{H} \mid \text{dist}(\mathbf{e}_h, \mathbf{e}_{l'}) < \min_{l \in A} \text{dist}(\mathbf{e}_h, \mathbf{e}_l)\}$ as the set of humans for whom l' provides improvement when added to A. Similarly define $\mathcal{H}_B^{l'}$ for set B.

For humans in $h \in \mathcal{H} \setminus \mathcal{H}_A^{l'}$, the closest agent in A remains closer than l', so $\min_{l \in A \cup \{l'\}} \text{dist}(\mathbf{e}_h, \mathbf{e}_l) = \min_{l \in A} \text{dist}(\mathbf{e}_h, \mathbf{e}_l)$, giving zero marginal improvement. Therefore:

$$f(A \cup \{l'\}) - f(A) = \frac{1}{|\mathcal{H}|} \sum_{h \in \mathcal{H}_A^{l'}} \left[\min_{l \in A} \text{dist}(\mathbf{e}_h, \mathbf{e}_l) - \text{dist}(\mathbf{e}_h, \mathbf{e}_{l'}) \right]$$

Since $A \subseteq B$, for any human h, we have $\min_{l \in A} \text{dist}(\mathbf{e}_h, \mathbf{e}_l) \geq \min_{l \in B} \text{dist}(\mathbf{e}_h, \mathbf{e}_l)$.

This implies that if l' provides improvement over B (i.e., $h \in \mathcal{H}_B^{l'}$), then l' also provides improvement over A (i.e., $h \in \mathcal{H}_A^{l'}$). Therefore, $\mathcal{H}_B^{l'} \subseteq \mathcal{H}_A^{l'}$.

For humans in $\mathcal{H}_B^{l'}$, the improvement when adding l' to A is at least as large as when adding l' to B, since $\min_{l \in A} \text{dist}(\mathbf{e}_h, \mathbf{e}_l) - \text{dist}(\mathbf{e}_h, \mathbf{e}_{l'}) \geq \min_{l \in B} \text{dist}(\mathbf{e}_h, \mathbf{e}_l) - \text{dist}(\mathbf{e}_h, \mathbf{e}_{l'})$.

Similarly, for B we have:

$$f(B \cup \{l'\}) - f(B) = \frac{1}{|\mathcal{H}|} \sum_{h \in \mathcal{H}_B^{l'}} \left[\min_{l \in B} \text{dist}(\mathbf{e}_h, \mathbf{e}_l) - \text{dist}(\mathbf{e}_h, \mathbf{e}_{l'}) \right]$$

Therefore:

$$\begin{aligned}
f(A \cup \{l'\}) - f(A) &= \frac{1}{|\mathcal{H}|} \sum_{h \in \mathcal{H}_A^{l'}} \left[\min_{l \in A} \text{dist}(\mathbf{e}_h, \mathbf{e}_l) - \text{dist}(\mathbf{e}_h, \mathbf{e}_{l'}) \right] \\
&= \frac{1}{|\mathcal{H}|} \sum_{h \in \mathcal{H}_B^{l'}} \left[\min_{l \in A} \text{dist}(\mathbf{e}_h, \mathbf{e}_l) - \text{dist}(\mathbf{e}_h, \mathbf{e}_{l'}) \right] \\
&\quad + \frac{1}{|\mathcal{H}|} \sum_{h \in \mathcal{H}_A^{l'} \setminus \mathcal{H}_B^{l'}} \left[\min_{l \in A} \text{dist}(\mathbf{e}_h, \mathbf{e}_l) - \text{dist}(\mathbf{e}_h, \mathbf{e}_{l'}) \right] \\
&\geq \frac{1}{|\mathcal{H}|} \sum_{h \in \mathcal{H}_B^{l'}} \left[\min_{l \in B} \text{dist}(\mathbf{e}_h, \mathbf{e}_l) - \text{dist}(\mathbf{e}_h, \mathbf{e}_{l'}) \right] \\
&= f(B \cup \{l'\}) - f(B)
\end{aligned}$$

This shows that f is submodular. □

E.3. Proof of Theorem 4.3

Theorem 4.3 (Performance Guarantee for REPPOP$_{\text{mapped-1}}$ and REPPOP$_{\text{mapped-2}}$). Let $\tilde{\mathcal{L}} = \{l_h | h \in \mathcal{H}\}$ be the proxy agent set where for each $h \in \mathcal{H}$, $l_h \in N_\rho(h)$, with $N_\rho(h)$ representing the ρ-neighborhood of h. Define the human coverage ratio $\gamma = f(L_{\mathcal{H}}^*)/f(L_{\mathcal{L}}^*) \in [0, 1]$, where $L_{\mathcal{H}}^*$ is the optimal subset from the human set and $L_{\mathcal{L}}^*$ is the optimal subset from the full agent set. If $L_{\tilde{\mathcal{L}}}^{\text{greedy}}$ is the subset of size M returned by the greedy algorithm on $\tilde{\mathcal{L}}$, then:

$$f(L_{\tilde{\mathcal{L}}}^{\text{greedy}}) \quad \geq \quad (1 - 1/e) \left(\gamma \cdot f(L_{\mathcal{L}}^*) - \rho \right),$$

where γ measures the cost of restricting the search space to humans (coverage quality) and ρ measures the cost of approximating each human by a proxy agent (imitation error). The value of γ is determined by how expressive the human set is relative to the full agent space, whereas ρ depends on the proxy construction strategy: uniform sampling in REPPOP$_{\text{mapped-1}}$ typically yields larger ρ, while greedy selection in REPPOP$_{\text{mapped-2}}$ achieves smaller ρ at the expense of higher computational cost.

Proof. In our context, $L_{\mathcal{H}}^* \subseteq \mathcal{H}$ represents the optimal subset of humans of size M that would be selected if we directly choose humans instead of agents. This is a theoretical construct for analysis purposes. In contrast, $L_{\mathcal{L}}^*$ is the optimal subset

of size M from the actual agent set $\mathcal{L}$. The human coverage ratio $\gamma = \frac{f(L^*_{\mathcal{H}})}{f(L^*_{\mathcal{L}})} \in [0, 1]$ measures how well selecting from the human set can approximate the optimal solution achievable using the full agent set. We start our proof by first decomposing the objective function as

$$f(L) = \sum_{h \in \mathcal{H}} f_h(L)$$

where

$$f_h(L) = \frac{1}{|\mathcal{H}|} \left[D_{\max} - \min_{l \in L} \text{dist}(\mathbf{e}_h, \mathbf{e}_l) \right]$$

For each human h in the optimal subset $L^*_{\mathcal{H}}$, consider its corresponding agent l_h in the proxy agent set $\tilde{\mathcal{L}}$. Let $L^*_{\tilde{\mathcal{H}}} = \{l_h | h \in L^*_{\mathcal{H}}\}$.

By definition, for each human $h \in \mathcal{H}$ and its corresponding proxy agent $l_h \in \tilde{\mathcal{L}}$, we have $\text{dist}(\mathbf{e}_h, \mathbf{e}_{l_h}) \leq \rho$ since l_h is in the ρ-neighborhood of h. Therefore:

$$|f_h(L^*_{\mathcal{H}}) - f_h(L^*_{\tilde{\mathcal{H}}})| \leq \frac{\rho}{|\mathcal{H}|}$$

The above inequality holds because the maximum distance deviation between any human and its proxy agent is at most ρ (by definition of the ρ-neighborhood).

Then:

$$|f(L^*_{\mathcal{H}}) - f(L^*_{\tilde{\mathcal{H}}})| = \left| \sum_{h \in \mathcal{H}} f_h(L^*_{\mathcal{H}}) - f_h(L^*_{\tilde{\mathcal{H}}}) \right| \leq \sum_{h \in \mathcal{H}} |f_h(L^*_{\mathcal{H}}) - f_h(L^*_{\tilde{\mathcal{H}}})| \leq \frac{\rho}{|\mathcal{H}|} \cdot |\mathcal{H}| = \rho$$

This gives us:

$$f(L^*_{\tilde{\mathcal{H}}}) \geq f(L^*_{\mathcal{H}}) - \rho \tag{1}$$

Since $L^*_{\tilde{\mathcal{L}}}$ is the optimal subset of size M from the proxy agent set $\tilde{\mathcal{L}}$, and $L^*_{\tilde{\mathcal{H}}}$ is a feasible solution of size M from $\tilde{\mathcal{L}}$, we have:

$$f(L^*_{\tilde{\mathcal{L}}}) \geq f(L^*_{\tilde{\mathcal{H}}}) \tag{2}$$

From the guarantees of the greedy algorithm for submodular function maximization (Nemhauser et al., 1978), if $L^{greedy}_{\tilde{\mathcal{L}}}$ is the subset of size M returned by the greedy algorithm on $\tilde{\mathcal{L}}$, we have:

$$f(L^{greedy}_{\tilde{\mathcal{L}}}) \geq (1 - \frac{1}{e}) f(L^*_{\tilde{\mathcal{L}}}) \tag{3}$$

Combining inequalities (1), (2), and (3), we get:

$$f(L^{greedy}_{\tilde{\mathcal{L}}}) \geq (1 - \frac{1}{e}) f(L^*_{\tilde{\mathcal{L}}}) \geq (1 - \frac{1}{e}) f(L^*_{\tilde{\mathcal{H}}}) \geq (1 - \frac{1}{e})(f(L^*_{\mathcal{H}}) - \rho)$$

Using the definition of human coverage ratio $\gamma = \frac{f(L^*_{\mathcal{H}})}{f(L^*_{\mathcal{L}})}$, we have $f(L^*_{\mathcal{H}}) = \gamma \cdot f(L^*_{\mathcal{L}})$. Substitution yields

$$f(L^{greedy}_{\tilde{\mathcal{L}}}) \geq (1 - \frac{1}{e})(\gamma \cdot f(L^*_{\mathcal{L}}) - \rho)$$

This gives us the performance guarantees for REPPOP$_{\text{mapped-1}}$ and REPPOP$_{\text{mapped-2}}$. □