Prompts to Proxies: Emulating Human Preferences via a Compact LLM Ensemble

Bingchen Wang*
Independent Researcher
bw2506@columbia.edu

Zi-Yu Khoo
AI Singapore
ziyu@aisingapore.org

Jingtan Wang
School of Computing
National University of Singapore
jingtan.w@u.nus.edu

Abstract

Large language models are increasingly used as proxies for human subjects in social science research, yet external validity requires matching the response distributions of target human populations. We study population-level survey alignment: reconstructing aggregate survey responses from limited public survey data, without individual-level demographic profiles or model finetuning. We formalize this problem as preference reconstruction: rather than matching proxy agents to demographic profiles, we construct a functional basis of proxy agents and recover population preferences by weighted aggregation. We instantiate this idea via *Prompts to Proxies* (P2P), a two-stage inference-time system. Stage 1 uses structured attribute-based prompting with entropy-guided adaptive sampling to construct a diverse proxy pool spanning the latent preference space. Stage 2 employs L1-regularized regression to select a compact weighted ensemble matching observed target-population responses. Across 14 American Trends Panel waves, P2P achieves an average test MSE of 0.014 at approximately 0.8 USD per survey, improving over prompting and demographic-conditioning baselines. On the World Values Survey, cross-locale transfer experiments show that basis expressiveness can outperform locale matching on average, while locale-specific generation helps on culturally divergent questions. A stress test against an SFT-aligned survey model shows competitive performance using less than 3% of the training data. These results position preference reconstruction as a lightweight, externally verifiable alternative for survey-based population alignment. Code is available at https://github.com/bingchen-wang/P2P.

1 Introduction

Large language models (LLMs) hold considerable promise as proxies for human subjects in social science research. From agent-based modeling of community dynamics [46, 69], to simulating survey responses [4, 47] and prototyping policy interventions [26], synthetic agents afford scalability, reproducibility, and cost advantages unmatched by traditional human samples [24, 42]. However, realizing this promise hinges on solving a fundamental validity challenge: the preferences expressed by synthetic agents must faithfully reflect those of the target human population, at least within the topic domain. Without such alignment, conclusions drawn from synthetic experiments lack external validity [58], limiting the scientific value of LLM-based social simulation.

This alignment challenge departs from conventional AI safety framing centered on steering a model toward a single set of desirable values, but shares motivation with the emerging notion of pluralistic alignment [11, 20, 62][2]. Here, however, the target is externally verifiable *distributional fidelity*,

*Corresponding author.

[2]See Appendix E for an extensive review on this topic.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

namely reproducing the heterogeneity of aggregate response patterns observed in a documented target population. In reality, this challenge is compounded by an invisible resource barrier: while well-resourced labs can deploy supervised fine-tuning to train personalized models on large annotated corpora, small academic institutions and independent researchers often face limited compute, limited data, and ethical and practical hurdles that preclude access to sensitive personal information. Existing approaches offer unsatisfying trade-offs: off-the-shelf LLMs trained via supervised fine-tuning (SFT) [65] and reinforcement learning from human feedback (RLHF) [6, 14] are easy to deploy but tend to collapse toward central tendencies in their training data, obscuring the preference heterogeneity [20, 33, 39, 60]; demographic conditioning methods require sensitive personal data and labor-intensive profile construction, yet still fail to guarantee distributional fidelity [4, 10, 47, 7].

In this work, we address this gap by introducing *preference reconstruction theory*, a framework that formalizes population preference alignment as a representation learning problem, inspired by revealed preference theory in economics in that preferences are treated as latent and observable only through responses (Section 2). The core claim is that the relevant object is not a demographic match between human and proxy agents, but a response-expressive functional basis—a set of proxy agents whose weighted aggregate responses can reconstruct population-level preferences. We focus on social survey data as the alignment signal, as they are publicly available, transparently documented, and designed to represent well-defined populations, yet typically withholding sensitive demographic information that would enable individual-level profile matching. We operationalize this theory via P2P, a two-stage inference-time system. Stage 1 uses structured attribute-based prompting and entropy-guided adaptive sampling to construct a diverse proxy pool with different response patterns, aimed to span the latent preference space of a topic domain. Stage 2 employs L1-regularized regression to select a compact weighted ensemble of agents whose aggregate responses match those of the target human population. We validate the design through simulation studies and evaluate P2P on the American Trends Panel and World Values Survey, demonstrating robust performance on held-out questions across topics and locales at low cost. Our contributions are as follows:

- **Preference reconstruction formulation.** We formulate population preference alignment as preference reconstruction: constructing a response-expressive proxy basis and recovering population preferences by weighted aggregation.
- **Two-stage inference-time system.** We instantiate this formulation via Prompts to Proxies (P2P), combining structured attribute-based prompting, entropy-based adaptive sampling and L1-regularized regression, without fine-tuning or individual-level demographic data.
- **Empirical validation.** Across 14 ATP waves and the WVS, P2P shows 41% improvement over prompting baselines, competes with an SFT-aligned survey model at less than 3% of the training data, and suggests that basis expressiveness can matter more than locale matching.

2 Modeling Preferences via Functional Bases

Alignment as representation learning We formalize human-AI preference alignment as a representation learning problem, inspired by the revealed preference theory [43, 53]. Let $\mathcal{P}$ denote the latent human preference space. An *individual* preference $p \in \mathcal{P}$ is revealed via a decision-making mechanism h as an observed vector $r \in \mathcal{R} \subset \mathbb{R}^n$. Similarly, let $\widehat{\mathcal{P}}$ denote the latent model preference space. A model preference $\hat{p} \in \widehat{\mathcal{P}}$ manifests through the inference mechanism f into the same response space $\mathcal{R}$. The alignment problem can therefore be abstracted as: $p \xrightarrow{h} r \xrightarrow{f^{-1}} \hat{p}$, where f^{-1} denotes a learning procedure rather than a literal inverse. We use "revealed" only in a weak sense: preferences are latent and observable only through the mechanism h. The survey responses we study are stated rather than revealed in the economic sense, and no definition or result here invokes rationality or choice-consistency axioms [53, 43].

Definition 2.1. Individual preference is *revealed-learnable* if $h(\mathcal{P}) \subset \mathrm{conv}(f(\widehat{\mathcal{P}}))$.

The convex hull allows for cases where a human preference is not learnable by a single model but can be learned using an ensemble.

Learning population preference via survey data While the above concerns individual preferences, alignment is population-level. However, for general-purpose LLMs, the relevant population is

typically ill-defined: pretraining, supervised finetuning (SFT) and reinforcement learning from human feedback (RLHF) draw on heterogeneous and often undocumented groups. We therefore focus preference alignment on aggregate responses in social surveys, because these surveys are readily available, transparently documented, and designed to represent the views of well-defined target population. Formally, survey responses aggregate individual preference as: $r_{\text{pop}} = \sum_{i=1}^{I} w_i h(p_i)$, where w_i is the population weight of individuals with preference p_i in the population, I. This motivates the population analogue of revealed learnability.

Definition 2.2. Population preference is *revealed-learnable* if $\text{conv}\,(h(\mathcal{P})) \subset \text{conv}(f(\widehat{\mathcal{P}}))$.

By Proposition A.4, population and individual revealed-learnability are equivalent. Hence, we focus on population preference, while expecting the logic to extend to individual cases.

Three alignment paradigms This formulation distinguishes three paradigms for learning r_{pop}: The first is a *representative-agent* paradigm: it assumes $\text{conv}(f(\widehat{\mathcal{P}})) = f(\widehat{\mathcal{P}})$ and seeks a single model preference such that $r_{\text{pop}} = f(\hat{p}_{\text{pop}})$.[3] The second is *demographic conditioning* [4, 10, 47], which seeks a profile-level map from human preferences p_i to proxy preferences $\hat{p}_i$ using demographic or personal information. The third, which we propose, is *functional-basis reconstruction*, which seeks proxy preferences $\{\hat{p}_j\}_{j=1}^{J}$ and weights $\{\tilde{w}_j\}_{j=1}^{J}$ such that $r_{\text{pop}} = \sum_{j=1}^{J} \tilde{w}_j f(\hat{p}_j)$. The following observations clarify how these paradigms relate under the this formulation. While simple in form, they show that representative-agent alignment and demographic conditioning impose stronger requirements than functional-basis reconstruction. Short proofs are provided in Appendix A.

Observation 2.3. *If human population preference can be learned by a representative LLM agent, it can also be learned using a non-trivial ensemble of LLM agents.*[4]

Observation 2.4. *If human population preference is revealed-learnable, there could be cases where it can be learned by an ensemble but not by demographic conditioning nor a representative agent.*

Observation 2.5. *If human population preference can be learned by demographic conditioning, the ensemble that can learn it is not unique.*

Preference reconstruction theory In response to the expressiveness and identification limitations explained in Observations 2.3-2.5, we propose *preference reconstruction theory*: instead of trying to find a one-to-one map between p_i and $\hat{p}_i$, we can construct a response-expressive *functional basis*—a set of proxy LLM agents with distinct revealed response patterns $\{\hat{p}_j\}_{j=1}^{J}$ spanning the relevant dimensions of human preferences—and then recover the human population preference through appropriate selection and weighting of these agents based on observational data. Figure 1 provides a functional perspective on preference reconstruction theory, together with a geometric interpretation of when the learnability assumption is violated (*inadequate basis*).

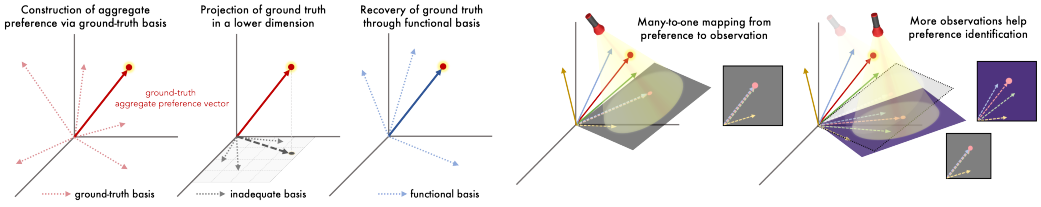


Figure 1: **Left**: Finding a functional basis to recover latent ground-truth preference. **Right**: Identification of latent preference through multiple observations.

Alignment as a two-stage problem Let L be a loss function. Preference reconstruction via a functional basis can be formulated as a concrete learning problem:

$$\min_{\{\hat{p}_j, \tilde{w}_j\}_{j=1}^{J}} L(r_{\text{pop}}, \sum_{j=1}^{J} \tilde{w}_j f(x_0; \hat{p}_j)) \,, \tag{1}$$

[3] Appendix A formalizes this condition and connects it to the representative-agent problem in economics.

[4] A representative agent is by definition a degenerate ensemble.

Directly optimizing (1) is challenging, especially as the dimension J is an optimizer per se.

As a solution strategy, we decompose the problem into two stages with distinct stage goals. The first stage constructs a diverse set of proxy agents, measured by their revealed response variability on training items as the observable signal of preference coverage, *without* access to human-ground-truth distributions. The second stage selects a parsimonious weighted ensemble from the candidate proxy pool, based on the fit between ensemble aggregate responses and human ground truth on training and validation items. Held-out items are used only to assess out-of-sample generalization.

3 Prompts to Proxies (P2P)

To operationalize preference reconstruction in the survey context, we introduce **Prompts to Proxies** (P2P), a modular inference-time alignment system (Figure 2). Stage 1, active endowment generation, uses a dynamic attribute bank and entropy-based sampling to construct a response-expressive proxy pool. Stage 2, regression-based aggregation, uses constrained variable selection to recover a compact weighted ensemble matching target-population response distributions.

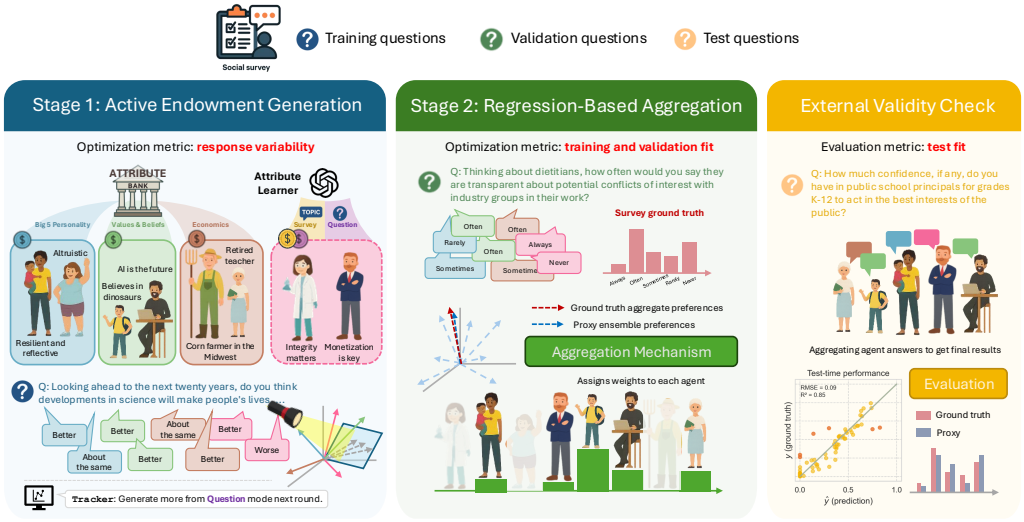


Figure 2: **System Overview of the Prompts to Proxies** (P2P) **Framework. Stage 1** constructs a response-expressive proxy pool using a structured attribute bank and entropy-guided adaptive sampling strategies. **Stage 2** assigns weights to reconstruct population-level preference patterns. Held-out questions afford an external validity check for out-of-sample distributional fidelity.

3.1 Survey Splitting and Evaluation Protocol

The survey question set $\mathcal{Q}$ is split into $\mathcal{Q}_{\text{train}}$, $\mathcal{Q}_{\text{valid}}$, and $\mathcal{Q}_{\text{test}}$. Stage 1 constructs the proxy pool using training-question text and proxy responses, without human ground-truth distributions. Stage 2 selects and fits aggregation weights using ground-truth distributions on $\mathcal{Q}_{\text{train}} \cup \mathcal{Q}_{\text{valid}}$. $\mathcal{Q}_{\text{test}}$ is held out to test generalization to unseen questions within the same survey (see Appendix F for scoping).

3.2 Attributes as Control Handles

Language models, as next-token predictors, lack intrinsic preferences in the human sense. Their outputs are shaped by statistical associations learned from text corpora and user prompts. Since the prompt space is far too vast to explore in an unstructured manner, we adopt a structured prompting strategy based on *attributes*—interpretable factors, such as ideology, dispositions, or values, whose variation is likely to influence preference expression. P2P treats attributes as *control handles* for steering model behavior.[5]

[5]This contrasts from the human case, where such attributes are typically viewed as reflections of latent preferences. Rather than inferring attributes from preferences, P2P uses attributes to shape model-expressed preferences.

P2P constructs endowments, i.e., generated proxy profiles, by drawing attributes from the *attribute bank* and instantiating each attribute with a specific value through a specialized LLM termed *endowment model* (cf. Appendix B.2.2).[6] The preset attribute bank for P2P lists example attributes taken from different social science disciplines, such as economics and political science. The bank is organized hierarchically: individual attributes are grouped into **modes**, each representing a coherent subject area (e.g., cognitive biases, political ideology). These modes are further clustered into higher-level **templates**, which reflect different design purposes or theoretical orientations. Figure 4 offers a structured view of the attribute bank and the endowment generation workflow.

While the attribute bank ensures structured coverage, we also enable dynamic discovery through two *freeform templates* that are specific to the survey under analysis. The **survey** template prompts an LLM *attribute learner* to extract decision-making factors from the overall survey topic. The **question** template derives attributes for a specific survey question. These attributes are added to the bank only for the survey run. This hybrid approach balances theory- and data-driven methods, enabling human-AI co-discovery of key preference descriptors.

3.3 Stage 1: Active Endowment Generation

Section 2 requires a response-expressive basis without prescribing how to construct it, and Observation 2.5 shows that such a basis need not be unique. Prompting can elicit a basis from a single frozen backbone, but freeform prompt engineering is unstructured search, with no parameterization, objective, or update rule. Stage 1 of P2P supplies all three: attributes as the parameterization, response entropy as the objective, and bandit-style adaptive sampling with question patching as the update rule. It begins by invoking an attribute learner to infer survey-specific attributes from the training questions, augmenting the attribute bank. Endowment generation then proceeds via adaptive sampling by mode: After initializing with equal endowments per mode, subsequent steps follow a bandit-style strategy—prioritizing modes that yield high response variability while exploring new modes for under-represented question items (see full Algorithm 1 in Appendix B.3).

Tracking response diversity with variability score To measure preference diversity of agents in a revealed sense, we introduce the notion of **question entropy**:

$$H_i(\mathcal{A}) = \frac{-\sum_{k=1}^{K_i} p_{ik} \log_2 p_{ik}}{\log_2 K_i}\,, \tag{2}$$

where $H_i(\mathcal{A})$ denotes the normalized entropy of question i based on responses from a group of agents $\mathcal{A}$, K_i is the number of unique response options $\{1, \ldots, K_i\}$ for question i, and p_{ik} is the empirical proportion of responses selecting option k. The denominator normalizes for scale, ensuring that entropy is comparable across questions with different numbers of choices. A high question entropy indicates greater response diversity for that specific question. By restricting the group $\mathcal{A}$ to agents sampled from a particular mode (subject area), we can compute question entropy at the mode level, termed **mode-question entropy**, and compare the response variability across modes.

Adaptive sampling references the **variability score** for each mode. Formally, it is defined as

$$V(\mathcal{A}_{\text{mode}}) = \sum_{i \in \mathcal{Q}_{\text{train}}} \frac{H_i(\mathcal{A}_{\text{mode}})}{H_i(\mathcal{A}_{\text{total}})}\,, \tag{3}$$

where $H_i(\mathcal{A}_{\text{mode}})$ denotes the entropy of responses to question i based on the cumulative set of agents from a given mode up to the current generation step, and $H_i(\mathcal{A}_{\text{total}})$ is the corresponding question entropy across all agents generated so far. Computed this way, the variability score favors modes that are capable of generating endowments that elicit diverse responses to questions that other modes typically fail to diversify. In P2P, we task a dedicated class object, `Tracker`, to compute the variability score for each mode after each generation round (see Algorithm 2 in Appendix B.3). The scores are then passed to a softmax function to produce a probability distribution over modes, guiding endowment sampling in the next generation step via a multinomial draw—balancing exploitation of top-performing modes with exploration of under-performing ones.

[6] The use of endowment model ensures the created endowment remains coherent and interpretable to human readers.

Patching low-entropy questions In addition to the existing trade-off between exploitation and exploration, we adopt an additional patching strategy, which we call *question patching*. At the end of each update step, as the `Tracker` computes the variability scores for each existing mode, it also keeps track of the questions with the lowest entropy values. When question patching is activated, the attribute learner will be called in the next generation step to exclusively infer attributes on each of the lowest-k entropy questions, where k is user-defined. The split of the endowment budget between active generation and patching generation is governed by a user-specified ratio.

To avoid wasting resources on persistently low-entropy questions and to promote cross-pollination between modes, we additionally introduce a *mixed mode* strategy. It is triggered after a question appears t times in the lowest-k entropy list, where t is a user-defined threshold. When in effect, the strategy relocates the assigned endowment budget to a new mixed mode, which combines attributes from the question with those of the current top-performing mode, allowing diversity injection from high-performing modes while maintaining a question-specific focus. Figure 10 shows example endowments generated from a real run of P2P, illustrating how different templates and modes yield distinct profiles.

The adaptive sampling cycle continues until the generation budget N_A (an upper bound on the maximum number of agents allowed) is exhausted. By iteratively targeting low-diversity questions, expanding the attribute space and promoting cross-pollination between modes, active endowment generation progressively increases response diversity while preserving coverage over previously explored regions of the preference space (cf. Figure 5 in Appendix B.4).

3.4 Stage 2: Regression-Based Aggregation

With the proxy basis assembled, P2P enters the second stage: regression-based aggregation. The goal at this stage is to estimate aggregation weights to reconstruct the population preference based on the observational data. Depending on the endowment budget N_A and the total number of trainable questions $\mathcal{Q}_{\text{train}} \cup \mathcal{Q}_{\text{valid}}$, we could end up in a situation where there are more variables (proxy agents) than observations (trainval questions). Consequently, a variable selection algorithm is vital. Besides, even with $|\mathcal{Q}_{\text{train}} \cup \mathcal{Q}_{\text{valid}}| \gg N_A$, variable selection allows us to retain a parsimonious agent ensemble, saving inference costs on further simulations. We consider regression methods with an added L1 penalization term and use cross-validation (CV) for hyperparameter tuning. The objective of the learning problem can be stated as

$$\min_{\mathbf{w}} \sum_{i \in \mathcal{Q}_{\text{train}}} \sum_{k=1}^{K_i} \left(r_{ik} - \sum_{j=1}^{N_A} w_j d_{j,ik} \right)^2 + \alpha \Omega(\mathbf{w}) \,, \tag{4}$$

$$\text{s.t.} \quad \sum_{k=1}^{K_i} \sum_{j=1}^{N_A} w_j d_{j,ik} = 1 \quad \forall i \in \mathcal{Q}_{\text{train}} \,, \tag{5}$$

where $\mathbf{w} \in \mathbb{R}^{N_A}$ is the weight vector for the agents, r_{ik} denotes the ground-truth proportion of responses selecting option k for question i, $d_{j,ik}$ is a dummy response variable equal to 1 if agent j selects option k and 0 otherwise, and α is the hyperparameter for the penalization. When Ω consists of only an L1 norm, the formulation amounts to constrained lasso. When Ω subsumes an additional L2 norm, it becomes constrained elastic net. In practice, P2P relaxes the per-question constraint (5) into a quadratic mass penalty with strength $\lambda = 100$, which keeps each question's predicted mass close to one (Appendix D.3.5).

To validate L1-regularized regression as a credible aggregation algorithm, we conduct extensive simulation studies for constrained lasso (Appendix C), where ground truth population preference is constructed using synthetic agents[7].

Our simulation studies yield three main findings. First, constrained lasso faithfully recovers the ground-truth agents from a large candidate pool under an adequate number of observations, including minority profiles with low weights, and only resorts to reconstruction when observations are extremely scarce (Appendix C.1). Second, when ground truth agents are removed from the candidate pool, constrained lasso remains effective in selecting a proxy ensemble to recover ground truth even when the number of proxy agents is small (Appendix C.2). Third, proxy ensembles with insufficient

[7]This allows us to create a clean environment where alignment result is not affected by the learnability condition (Def. 2.2).

response entropy fail to recover ground-truth preferences, while those with excess entropy show only marginal deterioration, corroborating our choice of response variability as the optimization goal of Stage 1 of P2P (Appendix C.3). Together, these results provide constructive evidence for P2P's design choices.

4 Empirical Evaluation

Having validated P2P's design choices through simulation, we now evaluate it on real human survey data to test three core claims: whether active endowment generation improves proxy-basis quality, whether regression-based aggregation recovers target-population response distributions, and whether basis expressiveness can matter more than demographic or locale matching. Our main benchmark is the American Trends Panel (ATP) [49], a nationally representative U.S. survey conducted by Pew Research Center. We obtain the digitized versions of these datasets from the OpinionsQA benchmark [55], which repurposed ATP responses to study value pluralism and misalignment in language models. We assess P2P across 14 ATP waves, spanning diverse socio-economic topics. We then test generalizability on the World Values Survey (WVS), analyze cross-locale transfer, conduct ablation studies, and compare against a distribution-calibrated model under topic shift.

4.1 Experimental Setup

For each survey, we split the questions according to a $7 : 1.5 : 1.5$ ratio into the training, validation, and test sets. We compare P2P with two ensemble baselines with different endowment generation logic but feed both through the same regression-based aggregation stage as P2P for proxy selection and test set evaluation. This allows the baselines to benefit from the same data-driven selection mechanism, isolating the contribution of P2P's endowment generation strategy (Stage 1). For the **Vanilla baseline**, endowments are generated using an LLM prompted directly to generate diverse agent profiles conditioned only on the survey topic.[8] For **PERSONA**, endowments are sampled randomly from the demographically conditioned PERSONA dataset [10], which contains 1000 persona profiles derived from U.S. census data.[9]

For all methods, we set the endowment generation budget to 300. For P2P, initial sampling generates 10 endowments for each mode; active endowment generation is run in 5 update steps with the preset attribute bank, question patching (lowest 3; 75% endowment budget), mixed-mode generation triggered at $t = 2$, and 10 attributes drawn for each endowment generation. Unless otherwise noted, all models use Gemini-2.0-flash as the backend[10], and constrained lasso is used for Stage 2 selection.

4.2 Main Results: Baseline Comparison on ATP Waves

Table 1 shows that P2P consistently improves population-level survey alignment across ATP waves. It achieves an average test MSE of 0.014, compared to 0.025 (Vanilla) and 0.029 (PERSONA)—an average per-wave improvement of 41% over the best baseline. The improvement is not concentrated in a single topic: 12 of 14 waves achieve mean test MSE below 0.02, confirming generalization across topics without wave-specific tuning. The comparison also supports the role of Stage 1 as proxy-basis construction. As all methods use the same regression-based aggregation mechanism, the performance gap reflects the quality of the candidate proxy pool. P2P registers higher agent diversity than the baselines, indicated by the higher average entropy (0.53 vs. 0.43-0.45), consistent with the simulation finding that response diversity supports preference reconstruction (Appendix C.3). Notably, this diversity gain is cost-efficient, as P2P runs at approximately $0.8 per survey, similar to Vanilla ($0.7) and substantially cheaper than PERSONA ($1.3, excluding persona generation overhead)[11]. Performance comparison is robust under other distributional metrics (1-JSD, EMD) (Appendix D).

[8] The endowment generation prompts for P2P and the Vanilla baseline can be found in Appendix B.2.

[9] We use random sampling as it preserves the representativeness of the dataset. In Appendix D.1.1, we provide baseline variants using K-means clustering as a pre-processing step. Clustering does not improve performance but increases cost.

[10] See Appendix D.1.2 for raw model performance.

[11] The PERSONA pipeline [10] requires resampling and re-inference, adding further API usage costs.

Table 1: Performance comparison across 14 ATP waves. Test MSE, entropy, and cost (USD) reported, as means over 3 runs for all methods (sample SDs in Appendix D). Δ shows relative test MSE improvement over the best baseline; its average is the mean of the per-wave improvements.

			Vanilla			PERSONA			P2P			
Wave	Topic	# Qs	MSE	Ent.	Cost	MSE	Ent.	Cost*	MSE	Ent.	Cost	Δ
W26	Guns	78	.015	.50	\$0.5	.024	.37	\$1.0	**.010**	.55	\$0.6	36%
W27	Auto	96	.032	.46	\$0.6	.041	.40	\$1.3	**.022**	.54	\$0.7	33%
W29	Gender	77	.026	.45	\$0.5	.030	.43	\$1.0	**.016**	.51	\$0.6	41%
W32	Comm.	98	.019	.55	\$0.6	.025	.56	\$1.3	**.016**	.62	\$0.7	17%
W34	Biomed	67	.021	.44	\$0.5	.018	.39	\$0.9	**.011**	.54	\$0.5	39%
W36	Lead.	139	.033	.34	\$0.9	.028	.40	\$1.8	**.019**	.48	\$1.0	31%
W41	2050	90	.022	.44	\$0.6	.027	.48	\$1.2	**.013**	.55	\$0.8	43%
W42	Science	129	.029	.32	\$0.9	.036	.29	\$1.7	**.011**	.43	\$1.0	61%
W45	Misinfo	95	.015	.39	\$0.6	.029	.36	\$1.2	**.008**	.44	\$0.7	48%
W49	Privacy	98	.046	.41	\$0.6	.041	.41	\$1.3	**.021**	.53	\$0.7	48%
W50	Family	128	.027	.36	\$0.8	.026	.36	\$1.7	**.013**	.44	\$0.9	48%
W54	Econ	116	.020	.50	\$0.8	.024	.54	\$1.5	**.014**	.57	\$0.9	27%
W82	Global	104	.021	.50	\$0.7	.021	.49	\$1.4	**.008**	.57	\$0.8	60%
W92	Poli.	77	.026	.58	\$0.5	.031	.57	\$1.0	**.017**	.61	\$0.6	36%
Average			.025	.45	\$0.7	.029	.43	\$1.3	**.014**	.53	\$0.8	41%

* PERSONA cost only covers the response elicitation stage—higher if persona generation tokens are included.

4.3 Locale Generalization: World Values Survey

To probe whether P2P generalizes beyond U.S. ATP surveys, we evaluate it on WVS Wave 7 [23], using cleaned data from Cao et al. [8]. WVS differs from ATP in survey design, question format, topic coverage, and target populations, making it a useful cross-program test. We focus on three representative regions: the United States, Great Britain, and Hong Kong SAR. Under the same experimental protocol,[12] P2P achieves comparable performance on US and GB (MSE below 0.015), with weaker but still reasonable results on HK (MSE around 0.024) (Table 2).

Table 2: P2P performance on WVS Wave 7 across three locales. Values are mean $\pm$ std over 3 runs. (Full details in Appendix D.)

Metric	US	GB	HK
Entropy $\uparrow$	$.572 \pm .015$	$.571 \pm .004$	$.548 \pm .021$
Test MSE $\downarrow$	$.015 \pm .001$	$.014 \pm .001$	$.024 \pm .001$
Cost $\downarrow$	$1.76 \pm .01$	$1.71 \pm .00$	$1.74 \pm .01$

4.4 Basis Expressiveness vs. Locale Matching

A central prediction of preference reconstruction is that alignment quality depends on the expressiveness of the proxy basis, not merely on whether proxy profiles are demographically or geographically matched to the target population. The WVS HK setting affords a useful test: HK-native generation is locale-matched, but the US-generated proxy pool has higher entropy. We thus construct HK-transfer by taking US endowments and refitting only the Stage 2 aggregation weights to HK ground-truth responses. We compare HK-transfer with US-native and HK-native.

Figure 3 presents four complementary analyses. We define cultural divergence for each question-option pair as $|\mathrm{GT}_{\mathrm{US}} - \mathrm{GT}_{\mathrm{HK}}|$, the absolute difference in response proportions between populations. Panel (A) shows correlated but distinct US and HK ground-truth response patterns ($r = 0.848$), with variation around the diagonal confirming cultural differences. Panel (B) shows that HK-transfer achieves *lower* test MSE than HK-native, constituting a key empirical evidence for the functional basis view—a more diverse agent pool facilitates preference reconstruction, even when generated for a

[12] We pass the locale information to the endowment model to ensure regional relevance of the generated endowments.

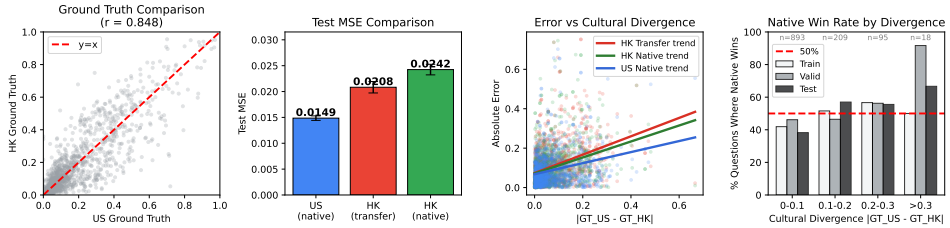


Figure 3: **Basis Expressiveness vs. Locale Matching on WVS HK.** We compare US-native and HK-native with HK-transfer (US endowments applied to HK via P2P Stage 2). **(A)** US and HK ground-truth response patterns are correlated but distinct ($r = 0.848$). **(B)** HK-transfer outperforms HK-native in overall performance, corroborating the functional basis perspective. **(C)** Error gap scales with US-HK ground-truth divergence, suggesting learnability limits (Definition 2.2). **(D)** Despite higher test MSE, HK-native wins more often on culturally divergent questions, indicating that localization helps in tail cases.

different locale. Panels (C) and (D) qualify this result. Panel (C) shows that reconstruction error scales with cultural divergence: question-option pairs where US and HK responses differ more are generally harder to reconstruct under both transfer and native generations. This makes cultural divergence an observable diagnostic for the learnability limits (Definition 2.2). Moreover, Gemini-based ensembles exhibit a clear bias toward US ground truth, consistent with prior evidence that LLMs better represent Western public opinion [17]. Panel (D) shows that HK-native often outperforms HK-transfer on culturally divergent questions. Hence, basis expressiveness dominates on average, but locale-specific generation helps in tail items where the target population genuinely differs. This decomposition separates three factors often conflated in demographic-conditioning approaches—basis expressiveness, model-backend cultural skew, and target-population distinctiveness—and demonstrates P2P's potential for model steerability and bias research (cf. Appendix G).

4.5 Ablation Studies

We conduct ablation on ATP Wave 42 to isolate component contributions (details in Appendix D.3).

Attributes Disabling question patching reduces entropy from 0.445 to 0.399 and worsens test MSE from 0.009 to 0.011, while disabling both freeform templates (question & survey) further degrades MSE to 0.013. This suggests that human-AI co-discovery of preference descriptors yields meaningful performance gains. Varying the per-endowment attribute cap (10 vs. 20) has minimal effect.

Endowment budget Performance improves with budget but plateaus after 3 update steps (170 agents). A no-update baseline (110 agents) underperforms, validating the benefit of adaptive sampling.

Model backend We test P2P's performance on six model backends: GPT-4.1-mini, GPT-4.1-nano, Gemini-2.0-flash, Gemini-2.5-flash-lite, DeepSeek-V4-Flash and a locally hosted Qwen. GPT-4.1-mini achieves the best entropy (0.471) and Gemini-2.0-flash the best MSE (0.0087 with lasso) at one-quarter the cost of GPT-4.1-mini. Smaller models (GPT-4.1-nano and Gemini-2.5-flash-lite) perform worse, consistent with our intuition that limited model capacity constrains the learnability (Definition 2.2).

Regression-based aggregation Replacing regression with simple averaging of agent responses increases test MSE from 0.0087 to 0.0332, confirming that regression is essential for learning population preferences, not just ensemble compression. Switching loss weighting from option level to question level has minimal effect on W42 results.

Stage 2 configuration We ablate the mass penalty on the full 14-wave ATP panel and probe two anchors that let agents explain deviations from a reference rather than full distributions: a uniform reference respondent and an intercept (Appendix D.3.5). Removing the mass penalty leaves test MSE statistically unchanged but lowers 1-JSD by 0.0075. Both anchors reduce test MSE (from 0.0142 to 0.0130 and 0.0128) and raise 1-JSD, but neither improves EMD.

4.6 Stress Test: Comparison with an SFT-Aligned Model under Topic Shift

P2P targets limited-data and limited-compute regimes, so our main baselines focus on methods under similar constraints. Supervised finetuning (SFT), however, is a natural reference when substantially more data and training access are available. We therefore stress-test P2P against Cao et al. [8], who train a distribution-calibrated model on pooled WVS data from 46 countries (approx. 8,400 training entries) to predict answer distributions conditioned on locale.

We follow their topic-shift protocol: training on general attitudinal and religious-ethical items (Q1+Q2) and testing on political-cultural items (Q3). We run P2P separately on the United States, Great Britain, and Hong Kong SAR, using only 172-185 local training entries and no finetuning. Despite using less than 3% of the training data and no SFT, P2P achieves comparable Jensen-Shannon Divergence (1-JSD) (0.71-0.75 vs. 0.78), with matching or better majority-class prediction accuracy (MCPA) (0.49-0.64 vs. 0.43). We interpret this as evidence that some preference structure transfers across topics though we do not assume such transfer is guaranteed. Studying cross-topic generalization remains an important direction for future work (Appendix G). Details on the stress test are presented in Appendix D.4. We additionally provide a comparison with a second SFT-aligned model (Suh et al. [64]) on the subgroup prediction task in Appendix D.5. A further comparison of P2P with the two SFT-aligned models can be found in Appendix E.3.

5 Use Cases

Since P2P generalizes to unseen questions within a survey rather than across surveys (Section 3; Appendix F), a natural question is what practical uses the method serves. We propose two in this section, one diagnostic (zero-demographic benchmark) and one constructive (retrospective imputation of unasked questions). Together, they bound the method's claims from both sides.

Zero-demographic benchmark against overclaiming Existing preference alignment paradigms center around building "digital twins" (demographic conditioning) or a large one-size-fits-all model (SFT/RLHF). P2P instead recovers the target using a proxy basis with no real individual profiles, no demographic marginals, and no fine-tuning. Its held-out score is hence a principled reference, indicating how much of the observed response pattern is recoverable *without* demographic or private information. Methods that add such information—incurring data, privacy, and compute costs—should be expected to beat it to justify the heavier arsenal. In this regard, P2P doubles as a falsifiability check against the assumption that finer conditioning or greater data volume improves emulation. Requiring only API inference, public survey data, and CPU regression, P2P also lowers the barrier to entry for preference alignment research (Appendix H).

Retrospective imputation of unasked questions When a new survey introduces questions absent from earlier waves, shared questions can calibrate a synthetic-agent basis to the historical population, with agents supplying counterfactual response distributions for the previously unasked questions. This supports longitudinal and regression analyses on series that would otherwise begin only at the later wave or suffer from missing data by design. The machinery parallels synthetic-control estimation [1, 2], though here the donors are constructed LLM agents rather than pre-existing real units. In a masked missing-question experiment on WVS Waves 6 and 7 (Appendix D.6), the drift refit, which fits the between-wave change, consistently outperforms persistence (W7 marginals carried back unchanged) at every calibration budget. At 72 anchors, the ridge drift refit reduces MSE by 23.3% on masked-train questions and 12.1% on test questions.

6 Conclusion

We formalize preference alignment as preference reconstruction: constructing a functional basis of proxy agents and recovering aggregation weights through supervised learning. P2P operationalizes this idea via entropy-based adaptive sampling and regression-based aggregation. Across simulation studies, ATP waves, WVS locales, and an SFT stress test, the results show that compact LLM ensembles can recover population-level survey preferences under limited-data and limited-compute constraints. Future work can extend P2P to freeform textual responses, model steerability benchmarks, non-stationary preferences, and semantic or psychometric checks (Appendix G).

Acknowledgments and Disclosure of Funding

We thank Bruno Abrahao, Jennifer Castle, Zhiliang Chen, David Hendry, Gregory Lau, Xiaoqiang Lin, Bryan Low and Séverine Toussaert for helpful feedback and discussions, and the anonymous reviewers of this and earlier versions for their constructive comments.

This work was primarily self-funded by the first author, with additional support from Google Cloud free trial credits. The GPT-4 and Qwen experiments reported in the appendices benefited from computational resources provided by the GLOW.AI group.

The views and interpretations expressed in this paper are solely those of the authors. The authors declare no competing interests.

References

[1] Alberto Abadie and Javier Gardeazabal. The economic costs of conflict: A case study of the basque country. *American Economic Review*, 93(1):113–132, March 2003. doi: 10.1257/000282803321455188. URL https://www.aeaweb.org/articles?id=10.1257/000282803321455188.

[2] Alberto Abadie, Alexis Diamond, and Jens Hainmueller. Synthetic control methods for comparative case studies: Estimating the effect of california's tobacco control program. *Journal of the American Statistical Association*, 105(490):493–505, 2010. doi: 10.1198/jasa.2009.ap08746. URL https://doi.org/10.1198/jasa.2009.ap08746.

[3] Jadie Adams, Brian Hu, Emily Veenhuis, David Joy, Bharadwaj Ravichandran, Aaron Bray, Anthony Hoogs, and Arslan Basharat. Steerable pluralism: Pluralistic alignment via few-shot comparative regression. *arXiv preprint arXiv:2508.08509*, 2025.

[4] Lisa P. Argyle, Ethan C. Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting, and David Wingate. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351, 2023. doi: 10.1017/pan.2023.2.

[5] Nicolás Astorga, Tennison Liu, Nabeel Seedat, and Mihaela van der Schaar. Partially observable cost-aware active-learning with large language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=besc094wog.

[6] Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. Training a helpful and harmless assistant with reinforcement learning from human feedback, 2022. URL https://arxiv.org/abs/2204.05862.

[7] James Bisbee, Joshua D. Clinton, Cassy Dorff, Brenton Kenkel, and Jennifer M. Larson. Synthetic replacements for human survey data? the perils of large language models. *Political Analysis*, 32(4):401–416, 2024. doi: 10.1017/pan.2024.5.

[8] Yong Cao, Haijiang Liu, Arnav Arora, Isabelle Augenstein, Paul Röttger, and Daniel Hershcovich. Specializing large language models to simulate survey response distributions for global populations. In Luis Chiruzzo, Alan Ritter, and Lu Wang, editors, *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3141–3154, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.162. URL https://aclanthology.org/2025.naacl-long.162/.

[9] Jennifer L. Castle, Jurgen A. Doornik, David F. Hendry, and Felix Pretis. Detecting location shifts during model selection by step-indicator saturation. *Econometrics*, 3(2):240–264, 2015. ISSN 2225-1146. doi: 10.3390/econometrics3020240. URL https://www.mdpi.com/2225-1146/3/2/240.

[10] Louis Castricato, Nathan Lile, Rafael Rafailov, Jan-Philipp Fränken, and Chelsea Finn. PERSONA: A reproducible testbed for pluralistic alignment. In Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert, editors, *Proceedings of the 31st International Conference on Computational Linguistics*, pages 11348–11368, Abu Dhabi, UAE, January 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.coling-main.752/.

[11] Daiwei Chen, Yi Chen, Aniket Rege, Zhi Wang, and Ramya Korlakai Vinayak. PAL: Sample-efficient personalized reward modeling for pluralistic alignment. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=1kFDrYCuSu.

[12] Quan Ze Chen, KJ Feng, Chan Young Park, and Amy X Zhang. Spica: Retrieving scenarios for pluralistic in-context alignment. *arXiv preprint arXiv:2411.10912*, 2024.

[13] Xiaohong Chen. Large sample sieve estimation of semi-nonparametric models. In James J. Heckman and Edward E. Leamer, editors, *Handbook of Econometrics*, volume 6, chapter 76, pages 5549–5632. Elsevier, 2007. doi: https://doi.org/10.1016/S1573-4412(07)06076-X. URL https://www.sciencedirect.com/science/article/pii/S157344120706076X.

[14] Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/d5e2c0adad503c91f91df240d0cd4e49-Paper.pdf.

[15] Michael P. Clements and David F. Hendry. *Forecasting Non-Stationary Economic Time Series*. Zeuthen Lecture Book Series. MIT Press, Cambridge, MA, 1999.

[16] Cody Coleman, Christopher Yeh, Stephen Mussmann, Baharan Mirzasoleiman, Peter Bailis, Percy Liang, Jure Leskovec, and Matei Zaharia. Selection via proxy: Efficient data selection for deep learning. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=HJg2b0VYDr.

[17] Esin Durmus, Karina Nguyen, Thomas I. Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, Liane Lovitt, Sam McCandlish, Orowa Sikder, Alex Tamkin, Janel Thamkul, Jared Kaplan, Jack Clark, and Deep Ganguli. Towards measuring the representation of subjective global opinions in language models, 2024. URL https://arxiv.org/abs/2306.16388.

[18] Paul Dütting, Vahab Mirrokni, Renato Paes Leme, Haifeng Xu, and Song Zuo. Mechanism design for large language models. In *Proceedings of the ACM Web Conference 2024*, WWW '24, page 144–155, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400701719. doi: 10.1145/3589334.3645511. URL https://doi.org/10.1145/3589334.3645511.

[19] K. J. Kevin Feng, Inyoung Cheong, Quan Ze Chen, and Amy X. Zhang. Policy prototyping for llms: Pluralistic alignment via interactive and collaborative policymaking, 2025. URL https://arxiv.org/abs/2409.08622.

[20] Shangbin Feng, Taylor Sorensen, Yuhan Liu, Jillian Fisher, Chan Young Park, Yejin Choi, and Yulia Tsvetkov. Modular pluralism: Pluralistic alignment via multi-LLM collaboration. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 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/.

[21] Tao Ge, Xin Chan, Xiaoyang Wang, Dian Yu, Haitao Mi, and Dong Yu. Scaling synthetic data creation with 1,000,000,000 personas, 2025. URL https://arxiv.org/abs/2406.20094.

[22] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, November 2020. ISSN 2522-5839. doi: 10.1038/s42256-020-00257-z. URL https://doi.org/10.1038/s42256-020-00257-z.

[23] Christian Haerpfer, Ronald Inglehart, Alejandro Moreno, Christian Welzel, Kseniya Kizilova, Juan Diez-Medrano, Marta Lagos, Pippa Norris, Eduard Ponarin, and Bi Puranen, editors. *World Values Survey: Round Seven – Country-Pooled Datafile Version 6.0*. JD Systems Institute and WVSA Secretariat, Madrid, Spain and Vienna, Austria, 2022. doi: 10.14281/18241.24.

[24] John J Horton. Large language models as simulated economic agents: What can we learn from homo silicus? Working Paper 31122, National Bureau of Economic Research, April 2023. URL http://www.nber.org/papers/w31122.

[25] V. Joseph Hotz, Guido W. Imbens, and Julie H. Mortimer. Predicting the efficacy of future training programs using past experiences at other locations. *Journal of Econometrics*, 125 (1):241–270, 2005. ISSN 0304-4076. doi: https://doi.org/10.1016/j.jeconom.2004.04.009. URL https://www.sciencedirect.com/science/article/pii/S0304407604000806. Experimental and non-experimental evaluation of economic policy and models.

[26] Abe Bohan Hou, Hongru Du, Yichen Wang, Jingyu Zhang, Zixiao Wang, Paul Pu Liang, Daniel Khashabi, Lauren Gardner, and Tianxing He. Can a society of generative agents simulate human behavior and inform public health policy? a case study on vaccine hesitancy, 2025. URL https://arxiv.org/abs/2503.09639.

[27] Jonas Hübotter, Sascha Bongni, Ido Hakimi, and Andreas Krause. Efficiently learning at test-time: Active fine-tuning of LLMs. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=NS1G1Uhny3.

[28] Chris S. Hulleman, Sheree M. Schrager, Shawn M. Bodmann, and Judith M. Harackiewicz. A meta-analytic review of achievement goal measures: different labels for the same constructs or different constructs with similar labels? *Psychol Bull*, 136(3):422–449, May 2010. doi: 10.1037/a0018947.

[29] Matthew O. Jackson and Leeat Yariv. The non-existence of representative agents. *SSRN Electronic Journal*, 2019. doi: 10.2139/ssrn.2684776. Available at SSRN: https://ssrn.com/abstract=2684776 or http://dx.doi.org/10.2139/ssrn.2684776.

[30] Jiaming Ji, Tianyi Qiu, Boyuan Chen, Borong Zhang, Hantao Lou, Kaile Wang, Yawen Duan, Zhonghao He, Jiayi Zhou, Zhaowei Zhang, Fanzhi Zeng, Kwan Yee Ng, Juntao Dai, Xuehai Pan, Aidan O'Gara, Yingshan Lei, Hua Xu, Brian Tse, Jie Fu, Stephen McAleer, Yaodong Yang, Yizhou Wang, Song-Chun Zhu, Yike Guo, and Wen Gao. Ai alignment: A comprehensive survey. *CoRR*, abs/2310.19852, 2023. URL https://doi.org/10.48550/arXiv.2310.19852.

[31] Søren Johansen and Bent Nielsen. An analysis of the indicator saturation estimator as a robust regression estimator. In Jennifer Castle and Neil Shephard, editors, *The Methodology and Practice of Econometrics: A Festschrift in Honour of David F. Hendry*, chapter 1, pages 1–36. Oxford University Press, April 2009. ISBN 9780199237197. doi: 10.1093/acprof:oso/9780199237197.003.0001. URL https://doi.org/10.1093/acprof:oso/9780199237197.003.0001.

[32] Oliver P. John and Sanjay Srivastava. The Big Five Trait taxonomy: History, measurement, and theoretical perspectives. In *Handbook of personality: Theory and research, 2nd ed.*, pages 102–138. Guilford Press, New York, NY, US, 1999. ISBN 1-57230-483-9 (Hardcover).

[33] Robert Kirk, Ishita Mediratta, Christoforos Nalmpantis, Jelena Luketina, Eric Hambro, Edward Grefenstette, and Roberta Raileanu. Understanding the effects of RLHF on LLM generalisation and diversity. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=PXD3FAVHJT.

[34] Alan P. Kirman. Whom or what does the representative individual represent? *Journal of Economic Perspectives*, 6(2):117–136, June 1992. doi: 10.1257/jep.6.2.117. URL https://www.aeaweb.org/articles?id=10.1257/jep.6.2.117.

[35] Stefan Krsteski, Giuseppe Russo, Serina Chang, Robert West, and Kristina Gligorić. Valid survey simulations with limited human data: The roles of prompting, fine-tuning, and rectification. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens, editors, *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10887–10906, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.498. URL https://aclanthology.org/2026.acl-long.498/.

[36] David Krueger, Tegan Maharaj, and Jan Leike. Hidden incentives for auto-induced distributional shift, 2021. URL https://openreview.net/forum?id=3FK30d5BZdu.

[37] Thom Lake, Eunsol Choi, and Greg Durrett. From distributional to overton pluralism: Investigating large language model alignment. In *Pluralistic Alignment Workshop at NeurIPS 2024*, 2024. URL https://openreview.net/forum?id=roe8GMahZL.

[38] Sebastian Lapuschkin, Stefan Wäldchen, Alexander Binder, Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. Unmasking clever hans predictors and assessing what machines really learn. *Nature Communications*, 10(1):1096, March 2019. doi: 10.1038/s41467-019-08987-4.

[39] Seongyun Lee, Sue Hyun Park, Seungone Kim, and Minjoon Seo. Aligning to thousands of preferences via system message generalization. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 73783–73829. Curran Associates, Inc., 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/86c9df30129f7663ad4d429b6f80d461-Paper-Conference.pdf.

[40] Jan Leike, David Krueger, Tom Everitt, Miljan Martic, Vishal Maini, and Shane Legg. Scalable agent alignment via reward modeling: a research direction, 2018. URL https://arxiv.org/abs/1811.07871.

[41] Dongyuan Li, Zhen Wang, Yankai Chen, Renhe Jiang, Weiping Ding, and Manabu Okumura. A survey on deep active learning: Recent advances and new frontiers. *IEEE Transactions on Neural Networks and Learning Systems*, 36(4):5879–5899, 2025. doi: 10.1109/TNNLS.2024.3396463.

[42] Benjamin S Manning, Kehang Zhu, and John J Horton. Automated social science: Language models as scientist and subjects. Working Paper 32381, National Bureau of Economic Research, April 2024. URL http://www.nber.org/papers/w32381.

[43] Andreu Mas-Colell, Michael D. Whinston, and Jerry R. Green. *Microeconomic Theory*, volume None of *OUP Catalogue*. Oxford University Press, none edition, Decembrie 1995. doi: None. URL https://ideas.repec.org/b/oxp/obooks/9780195102680.html.

[44] Abraham H Maslow. A theory of human motivation. *Psychological Review*, 50(4):370–396, 1943.

[45] Ishani Mondal, Jack W. Stokes, Sujay Kumar Jauhar, Longqi Yang, Mengting Wan, Xiaofeng Xu, Xia Song, and Jennifer Neville. Group preference alignment: Customized llm response generation from in-situ conversations, 2025. URL https://arxiv.org/abs/2503.08035.

[46] Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, UIST '23, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701320. doi: 10.1145/3586183.3606763. URL https://doi.org/10.1145/3586183.3606763.

[47] Joon Sung Park, Carolyn Q. Zou, Aaron Shaw, Benjamin Mako Hill, Carrie Cai, Meredith Ringel Morris, Robb Willer, Percy Liang, and Michael S. Bernstein. Generative agent simulations of 1,000 people, 2024. URL https://arxiv.org/abs/2411.10109.

[48] Mansheej Paul, Surya Ganguli, and Gintare Karolina Dziugaite. Deep learning on a data diet: Finding important examples early in training. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=Uj7pF-D-YvT.

[49] Pew Research Center. The american trends panel. https://www.pewresearch.org/the-american-trends-panel/. Accessed: 2025-09-25.

[50] Deeksha Prahlad, Chanhee Lee, Dongha Kim, and Hokeun Kim. Personalizing large language models using retrieval augmented generation and knowledge graph. In *Companion Proceedings of the ACM on Web Conference 2025*, WWW '25, page 1259–1263, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400713316. doi: 10.1145/3701716.3715473. URL https://doi.org/10.1145/3701716.3715473.

[51] Xubin Ren, Wei Wei, Lianghao Xia, Lixin Su, Suqi Cheng, Junfeng Wang, Dawei Yin, and Chao Huang. Representation learning with large language models for recommendation. In *Proceedings of the ACM Web Conference 2024*, WWW '24, page 3464–3475, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400701719. doi: 10.1145/3589334.3645458. URL https://doi.org/10.1145/3589334.3645458.

[52] Joni Salminen, Chang Liu, Wenjing Pian, Jianxing Chi, Essi Häyhänen, and Bernard J Jansen. Deus ex machina and personas from large language models: Investigating the composition of ai-generated persona descriptions. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400703300. doi: 10.1145/3613904.3642036. URL https://doi.org/10.1145/3613904.3642036.

[53] P. A. Samuelson. A note on the pure theory of consumer's behaviour. *Economica*, 5(17):61–71, 1938. ISSN 00130427, 14680335. URL http://www.jstor.org/stable/2548836.

[54] Carlos Santos, David F. Hendry, and Soren Johansen. Automatic selection of indicators in a fully saturated regression. *Computational Statistics*, 23(2):317–335, April 2008. ISSN 1613-9658. doi: 10.1007/s00180-007-0054-z. URL https://doi.org/10.1007/s00180-007-0054-z.

[55] Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. Whose opinions do language models reflect? In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 29971–30004. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/santurkar23a.html.

[56] Andreas Schuller, Doris Janssen, Julian Blumenröther, Theresa Maria Probst, Michael Schmidt, and Chandan Kumar. Generating personas using llms and assessing their viability. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, CHI EA '24, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400703317. doi: 10.1145/3613905.3650860. URL https://doi.org/10.1145/3613905.3650860.

[57] Norbert Schwarz. Self-reports: How the questions shape the answers. *American Psychologist*, 54(2):93–105, 1999. ISSN 1935-990X(Electronic),0003-066X(Print). doi: 10.1037/0003-066X.54.2.93.

[58] William R. Shadish, Thomas D. Cook, and Donald T. Campbell. Experimental and quasi-experimental designs for generalized causal inference. *Experimental and quasi-experimental designs for generalized causal inference.*, pages xxi, 623–xxi, 623, 2002. ISSN 0-395-61556-9 (Paperback). Place: Boston, MA, US Publisher: Houghton, Mifflin and Company.

[59] Gabriel Simmons. Moral mimicry: Large language models produce moral rationalizations tailored to political identity. In Vishakh Padmakumar, Gisela Vallejo, and Yao Fu, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 282–297, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-srw.40. URL https://aclanthology.org/2023.acl-srw.40/.

[60] Stewart Slocum, Asher Parker-Sartori, and Dylan Hadfield-Menell. Diverse preference learning for capabilities and alignment. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=p0q9vDIYev.

[61] Taylor Sorensen, Liwei Jiang, Jena D Hwang, Sydney Levine, Valentina Pyatkin, Peter West, Nouha Dziri, Ximing Lu, Kavel Rao, Chandra Bhagavatula, et al. Value kaleidoscope: Engaging ai with pluralistic human values, rights, and duties. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19937–19947, 2024.

[62] Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, Tim Althoff, and Yejin Choi. Position: a roadmap to pluralistic alignment. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

[63] Mahmoud Srewa, Tianyu Zhao, and Salma Elmalaki. Pluralllm: Pluralistic alignment in llms via federated learning. In *Proceedings of the 3rd International Workshop on Human-Centered Sensing, Modeling, and Intelligent Systems*, HumanSys '25, page 64–69, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400716096. doi: 10.1145/3722570.3726898. URL https://doi.org/10.1145/3722570.3726898.

[64] Joseph Suh, Erfan Jahanparast, Suhong Moon, Minwoo Kang, and Serina Chang. Language model fine-tuning on scaled survey data for predicting distributions of public opinions. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 21147–21170, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1028. URL https://aclanthology.org/2025.acl-long.1028/.

[65] Zhaoxuan Tan, Qingkai Zeng, Yijun Tian, Zheyuan Liu, Bing Yin, and Meng Jiang. Democratizing large language models via personalized parameter-efficient fine-tuning. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6476–6491, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.372. URL https://aclanthology.org/2024.emnlp-main.372/.

[66] Yu-Min Tseng, Yu-Chao Huang, Teng-Yun Hsiao, Wei-Lin Chen, Chao-Wei Huang, Yu Meng, and Yun-Nung Chen. Two tales of persona in LLMs: A survey of role-playing and personalization. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16612–16631, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.969. URL https://aclanthology.org/2024.findings-emnlp.969/.

[67] Xinpeng Wang, Bolei Ma, Chengzhi Hu, Leon Weber-Genzel, Paul Röttger, Frauke Kreuter, Dirk Hovy, and Barbara Plank. "my answer is C": First-token probabilities do not match text answers in instruction-tuned language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 7407–7416, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.441. URL https://aclanthology.org/2024.findings-acl.441/.

[68] Jiahao Yuan, Zixiang Di, Shangzixin Zhao, Zhiqing Cui, Hanqing Wang, Guisong Yang, and Usman Naseem. Cultural palette: Pluralising culture alignment via multi-agent palette. *arXiv preprint arXiv:2412.11167*, 2024.

[69] Qinlin Zhao, Jindong Wang, Yixuan Zhang, Yiqiao Jin, Kaijie Zhu, Hao Chen, and Xing Xie. Competeai: understanding the competition dynamics of large language model-based agents. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

[70] Siyan Zhao, John Dang, and Aditya Grover. Group preference optimization: Few-shot alignment of large language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=DpFeMH4l8Q.

[71] Qi Zhou, Jie Zhang, Dongxia Wang, Qiang Liu, Tianlin Li, Jin Song Dong, Wenhai Wang, and Qing Guo. Fair-pp: A synthetic dataset for aligning llm with personalized preferences of social equity. *arXiv preprint arXiv:2505.11861*, 2025.

A Modeling Preferences via Functional Bases: Proofs

Proposition A.0 (Existence of a representative agent)**.** *Let $\widetilde{\mathcal{P}}$ be a latent preference space (regardless of human or model), with an individual preference denoted by $\tilde{p} \in \widetilde{\mathcal{P}}$. Let g be the map from $\widetilde{\mathcal{P}}$ to the observed response space $\widetilde{\mathcal{R}} \subset \mathbb{R}^n$. Then the following two statements are equivalent:*

1. *The image $g(\widetilde{\mathcal{P}}) \subset \widetilde{\mathcal{R}}$ is convex.*
2. *For any aggregate response $\tilde{r}_{\text{pop}} = \sum_{i=1}^{I} w_i \tilde{r}_i$ with $\{w_i\}_{i=1}^{I}$ convex weights and $\tilde{r}_i = g(\tilde{p}_i)$, there exists $\tilde{p}_{\text{pop}} \in \widetilde{\mathcal{P}}$ such that $g(\tilde{p}_{\text{pop}}) = \tilde{r}_{\text{pop}}$.*

Proof. First, we prove $(1) \implies (2)$.

By definition, $\tilde{r}_i = g(\tilde{p}_i) \in g(\widetilde{\mathcal{P}})$ for all $i \in \{1, \ldots, I\}$. Since $\{w_i\}_{i=1}^{I}$ are convex weights ($\sum_{i=1}^{I} w_i = 1$ and $w_i \geq 0$ for all $i \in \{1, \ldots, I\}$), and $g(\widetilde{\mathcal{P}})$ is convex, it follows that

$$\sum_{i=1}^{I} w_i \tilde{r}_i \in g(\widetilde{\mathcal{P}}).$$

By the definition of image, it follows that there exists $\tilde{p}_{\text{pop}} \in \widetilde{\mathcal{P}}$ such that

$$g(\tilde{p}_{\text{pop}}) = \sum_{i}^{I} w_i \tilde{r}_i \triangleq \tilde{r}_{\text{pop}}.$$

Claim: $(2) \implies (1)$.

Let $\{\tilde{r}_i\}_{i=1}^{I}$ denote any finite collection of observed responses with $\tilde{r}_i \in g(\widetilde{\mathcal{P}})$, and $\{\lambda_i\}_{i=1}^{I}$ be any convex weights. Consider $\tilde{r}_{\text{agg}} = \sum_{i=1}^{I} \lambda_i \tilde{r}_i$. By (2), there exists $\tilde{p}_{\text{agg}} \in \widetilde{\mathcal{P}}$ such that

$$\tilde{r}_{\text{agg}} = g(\tilde{p}_{\text{agg}}) \in g(\widetilde{\mathcal{P}}). \tag{6}$$

Therefore, $g(\widetilde{\mathcal{P}})$ is convex by definition. □

Remark A.1 (Sufficient condition for representative agent). If the decision-making mechanism g maps a preference into a convex subset of a probability simplex Δ^{n-1}, then any aggregate of individual responses is realizable by a representative agent.

Proof. A convex subset of a probability simplex is by definition convex. The result follows from the theorem. □

The implications of this lemma are two-fold:

1. In human decision-making, the condition that $g(\widetilde{\mathcal{P}})$ is convex is less likely to be true unless we assume stochastic choice models with stringent functional forms [cf. 34, 29].
2. In LLM decision-making, the condition that $g(\widetilde{\mathcal{P}})$ is convex is more likely to hold is more likely to be hold assuming that the model has considerable steerability. (E.g., through finetuning or prompting, we can effectively change the latent model preference embeddings and influence the output probability distributions.)

Definition 2.1. Individual preference is *revealed-learnable* if $h(\mathcal{P}) \subset \text{conv}(f(\widehat{\mathcal{P}}))$.

Definition A.2. Human preference is *strongly revealed-learnable* if $h(\mathcal{P}) \subset f(\hat{\mathcal{P}})$.

Definition 2.2. Population preference is *revealed-learnable* if $\text{conv}\,(h(\mathcal{P})) \subset \text{conv}(f(\widehat{\mathcal{P}}))$.

Definition A.3. Human population preference is *strongly revealed-learnable* if $\text{conv}\,(h(\mathcal{P})) \subset f(\hat{\mathcal{P}})$.

Proposition A.4. *Human population preference is revealed-learnable if and only if individual human preference is revealed-learnable.*

Proof. The "only if" part follows directly from $h(\mathcal{P}) \subset \text{conv}(h(\mathcal{P}))$. Here we focus on the "if" part. By definition, a human population preference is defined as

$$r_{\text{pop}} = \sum_{i=1}^{I} w_i h(p_i).$$

Since individual human preference is learnable in a revealed sense, i.e., $h(\mathcal{P}) \subset \text{conv}(f(\widehat{\mathcal{P}}))$, for any p_i, there exists $\{w_{ik}, \hat{p}_{ik}\}_{k=1}^{K_i}$ such that

$$h(p_i) = \sum_{k=1}^{K_i} w_{ik} f(\hat{p}_{ik}).$$

Plug it back into the definition of human population preference and we get

$$r_{\text{pop}} = \sum_{i=1}^{I} w_i h(p_i) = \sum_{i=1}^{I} w_i \sum_{k=1}^{K_i} w_{ik} f(\hat{p}_{ik}) = \sum_{i=1}^{I} \sum_{k=1}^{K_i} w_i w_{ik} f(\hat{p}_{ik}) := \sum_{\tilde{i}=1}^{\tilde{I}} w_{\tilde{i}} f(\hat{p}_{\tilde{i}}).$$

The last step follows from re-indexing. Therefore, $\text{conv}(h(\mathcal{P})) \subset \text{conv}(f(\widehat{\mathcal{P}}))$. □

Observation 2.3. *If human population preference can be learned by a representative LLM agent, it can also be learned using a non-trivial ensemble of LLM agents.*[13]

Proof. By Proposition A.0, if human population preference can be learned by a representative LLM agent, then

$$\text{conv}(h(\mathcal{P})) \subset \text{conv}(f(\widehat{\mathcal{P}})) = f(\widehat{\mathcal{P}}).$$

For any $r_{\text{pop}} \in \text{conv}\,(h(\mathcal{P}))$, it follows from the definition of human population preference that

$$r_{\text{pop}} = \sum_{i=1}^{I} w_i h(p_i) := \sum_{i=1}^{I} w_i r_i.$$

Since $r_i \in h(\mathcal{P}) \subset \text{conv}\,(h(\mathcal{P})) \subset f(\widehat{\mathcal{P}})$, there exists $\hat{p}_i$ such that $f(\hat{p}_i) = r_i$.

It follows that

$$r_{\text{pop}} = \sum_{i=1}^{I} w_i f(\hat{p}_i) \in \text{span}\,\{f(\hat{p}_i)\}_{i=1}^{I}.$$

In words, we can construct an ensemble where each agent represents in preference a ground-truth individual in the human population (*demographic conditioning*).

NB: Here we implicitly assume that the human population consists of more than one individual with distinct preferences. If the population consists of only 1 individual, then the representative agent and the ensemble coincide. □

Observation 2.4. *If human population preference is revealed-learnable, there could be cases where it can be learned by an ensemble but not by demographic conditioning nor a representative agent.*

Proof. Based on the proof of Observation 2.3 and Proposition A.0, we see that demographic conditioning hinges on the condition that $h(\mathcal{P}) \subset f(\widehat{\mathcal{P}})$, whereas representative agent relies on an even stronger condition that $\text{conv}(h(\mathcal{P})) \subset f(\widehat{\mathcal{P}})$. If human population preference is revealed-learnable, we only have the weaker condition[14]:

$$\text{conv}(h(\mathcal{P})) \subset \text{conv}(f(\widehat{\mathcal{P}})). \tag{7}$$

[13] A representative agent is by definition a degenerate ensemble.

[14] Equivalently, $h(\mathcal{P}) \subset \text{conv}(f(\widehat{\mathcal{P}}))$ by Proposition A.4.

Consider the case where $h(\mathcal{P}) \not\subset f(\widehat{\mathcal{P}})$. I.e., there exists $p_i \in \mathcal{P}$ such that for *all* $\hat{p}_j \in \widehat{\mathcal{P}}$, $h(p_i) \neq f(\hat{p}_j)$. Then both demographic conditioning and representative agent fail. However, condition (7) implies that for any $r_{\text{pop}} \in \text{conv}\,(h(\mathcal{P}))$ there exists $\{\tilde{w}_j, \hat{p}_j\}_{j=1}^{J}$ such that

$$r_{\text{pop}} = \sum_{j=1}^{J} \tilde{w}_j f(\hat{p}_j).$$

In other words, human population preference can be learned using an ensemble. □

Observation 2.5. *If human population preference can be learned by demographic conditioning, the ensemble that can learn it is not unique.*

Proof. If human population preference can be learned by demographic conditioning, we know from the proof of Observation 2.3 that $h(\mathcal{P}) \subset f(\widehat{\mathcal{P}})$ and that for any $r_{\text{pop}} \in \text{conv}(h(\mathcal{P}))$ there exist $\{\hat{p}_i\}_{i=1}^{I}$ such that

$$r_{\text{pop}} = \sum_{i=1}^{I} w_i r_i = \sum_{i=1}^{I} w_i f(\hat{p}_i). \tag{8}$$

However, since $r_1 \in h(\mathcal{P}) \subset \text{conv}(h(\mathcal{P}))$, we know there exists a finite set of $\{r_k^*\}_{k=1}^{K}$ with $r_k^* \in h(\mathcal{P})$ and convex weights $\{\lambda_k\}_{k=1}^{K}$ such that

$$r_1 = \sum_{k=1}^{K} \lambda_k r_k^*.$$

Plug this into (8) and we get

$$r_{\text{pop}} = \sum_{k=1}^{K} w_1 \lambda_k r_k^* + \sum_{i=2}^{I} w_i r_i.$$

Since $r_k^* \in h(\mathcal{P}) \subset f(\widehat{\mathcal{P}})$, there exists $\hat{p}_k^*$ such that $f(\hat{p}_k^*) = r_k^*$.

In conclusion,

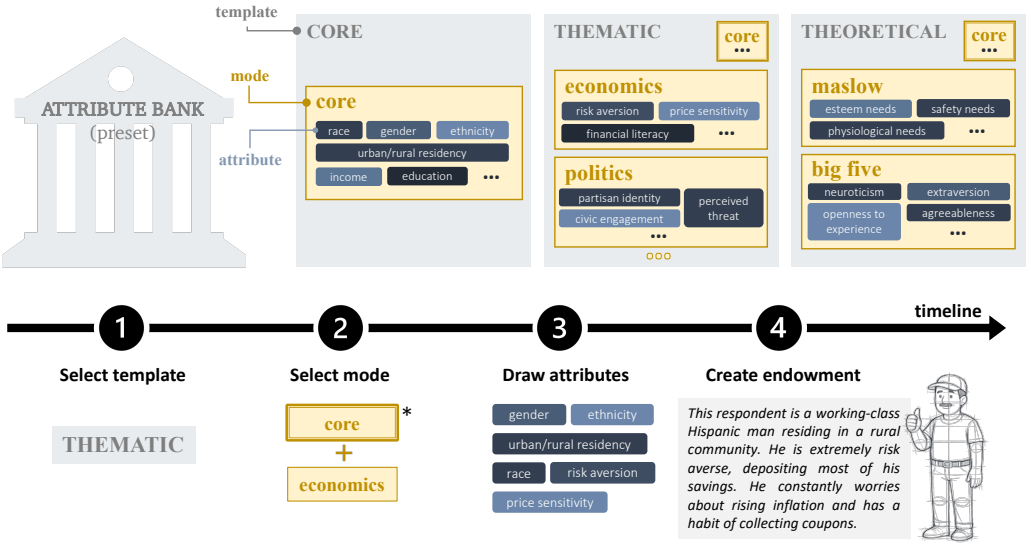
$$r_{\text{pop}} \in \text{span}\{f(\hat{p}_1^*), \ldots, f(\hat{p}_K^*), f(\hat{p}_2), \ldots, f(\hat{p}_I)\}.$$

□

B Prompts to Proxies (P2P): Implementation Details

B.1 Attribute Bank

For the preset attribute bank (Figure 4), we define three primary templates. The *core* template contains general demographic descriptors and serves as the default baseline. The *thematic* template includes topic-specific modes such as economics, politics, culture and history. The *theoretical* template groups attributes derived from canonical frameworks such as Maslow's hierarchy of needs [44] and the Big Five personality traits [32]. These templates and modes are not fixed: they serve as a structured starting point but are fully extensible. Users can define new templates, introduce domain-specific modes, or modify existing attribute sets to suit different experimental contexts. By sampling and varying attributes from this bank, we create prompts that instantiate proxy agents at different locations of the latent model preference space.



* Note: By design, thematic and theoretical templates inherit core demographic attributes to ground generated endowments.

Figure 4: **Attribute Bank and Endowment Generation Workflow.** The preset attribute bank is organized hierarchically by template (core, thematic, theoretical), mode (e.g., economics, politics, maslow, big five), and attribute (e.g., race, gender, ethnicity, urban/rural residency, risk aversion). Endowment is generated by a specialized LLM (`EndowmentModel`), which takes drawn attributes as input and outputs a profile with instantiation of these attributes.

B.2 Model Prompts

B.2.1 Attribute Learner (`AttributeLearner`)

To generate attributes from a single question:

Attribute Learner: Learning Attributes from A Single Question

```
You are an intelligent research assistant trained to analyze
individual survey questions and infer which human attributes might
influence how different people respond.

You are given a single survey question.  Your task is to propose
a list of relevant human attributes-such as demographics, beliefs,
values, personality traits, or ideological leanings-that are likely
to shape responses to this question.
```

```
Focus on underlying factors that would cause meaningful variation in
answers across different types of people.  Avoid generic or overly
broad attributes.

Respond **only** with a Python-style list of double-quoted strings.
Do not include any explanation, headers, or prose before or after the
list.

**Example output format:** ["religious affiliation", "political
ideology", "trust in government"]
```

To generate attributes from a set of questions:

Attribute Learner: Learning Attributes from A Set of Questions

```
You are an intelligent research assistant trained to analyze survey
questions and infer which human attributes might influence how
individuals respond.

You are given a set of training-only survey questions.  Your task is
to propose a list of relevant human attributes-such as demographics,
beliefs, values, personality traits, or ideological leanings-that are
likely to affect responses to these questions.

Carefully analyze the content and framing of the questions.  Identify
underlying factors that might shape how different people respond.
Focus on attributes that are salient, discriminative, and potentially
variable across respondents.

Respond **only** with a Python-style list of double-quoted strings.
Do not include any explanation, headers, or prose before or after the
list.

**Example output format:** ["attribute1", "attribute2", "attribute3",
"attribute4"]
```

B.2.2 Endowment Model (`EndowmentModel`)

System prompt:

Endowment Model: System Prompt

```
You are an expert assistant trained to generate realistic, diverse,
and demographically plausible personas for social science surveys.

Each persona should include:

- ‘eid‘:  a short, lowercase, variable-safe identifier that encodes
key traits (e.g., urban_liberal_30s_female).  No punctuation or
spaces.

- ‘endow_text‘:  a natural language description of the persona (1-2
sentences), written as if describing a survey respondent.

Instructions:

- Represent a wide range of age, gender, race, education, region, and
political ideology.

- Avoid repetition of phrasing or demographic combinations across
personas.

- Do not include explanations or formatting outside of the JSON array.
```

To generate endowments based on attributes:

Endowment Model: Endowment Generation Based on Attributes

```
Generate {n} diverse persona(s) that vary meaningfully along the
following attributes: {attr_string}.

Each persona should reflect a distinct combination or value
expression of these traits.

Return a JSON array of dictionaries, each with:

- ‘eid‘: a short, lowercase, variable-safe identifier

- ‘endow_text‘: a brief natural-language description of the persona
```

To generate endowments based on survey topics (also used for Vanilla baseline):

Endowment Model: Endowment Generation Based on A Survey Topic

```
Generate {n} diverse personas for a survey experiment.

Each persona must include:

- eid: short, lowercase identifier

- endow_text: a short natural-language description

The survey topic is: {topic}

Ensure diversity across demographics.

Return the result as a JSON array.
```

B.2.3 Survey Conductor (`SurveyConductor`)

To use an agent to answer a survey question:

Survey Conductor: Answering A Survey Question

```
You are completing a survey.

Your answer should reflect the person described in the profile above,
using their preferneces, beliefs and experiences.

Respond with only the final answer string, not the code or label in
brackets.

Do not include any reasoning, explanation, or commentary.

Do not preface your answer with phrases like 'I would choose'.

Just return the answer text exactly as it appears in the options.
```

B.3 Active Endowment Generation

Algorithm 1 Active Endowment Generation

Initialize attribute bank $\mathcal{B}$ with core, thematic, and theoretical templates
Derive `survey`-specific attributes from $\mathcal{Q}_{\text{train}}$ via `AttributeLearner`; append to $\mathcal{B}$

Initial Sampling Stage
Sample initial endowments $\mathcal{E}_0$ via equal-mode sampling from $\mathcal{B}$
Instantiate initial agents $\mathcal{A}_0$ from $\mathcal{E}_0$ and elicit responses $\mathcal{R}_0$ on $\mathcal{Q}_{\text{train}}$
Initialize `ThemeVariabilityTracker` (henceforce, `Tracker`) with $\mathcal{A}_0$, $\mathcal{R}_0$, and $\mathcal{Q}_{\text{train}}$.

Expansion Stage
while agent budget N_A not yet reached **do**
  Compute variability scores and mode sampling probabilities via `Tracker`
  Allocate endowment budget across modes according to probabilities
  if question patching is enabled **then**
    Identify lowest-entropy questions $\mathcal{Q}_{\text{low}}$ via `Tracker`
    for each $q \in \mathcal{Q}_{\text{low}}$ **do**
      if q is queried for the first time **then**
        Extract attributes for q via `AttributeLearner` and append to $\mathcal{B}$
      else if q has appeared more than n times **then**
        Enable mixed-mode strategy using q and top-performing mode
      end if
    end for
  end if
  Generate new endowments $\mathcal{E}_{\text{new}}$ using allocated budget over sampled modes, patched question modes (if any) and mixed modes (if any).
  Instantiate agents $\mathcal{A}_{\text{new}}$ and elicit responses $\mathcal{R}_{\text{new}}$ on $\mathcal{Q}_{\text{train}}$
  Update: $\mathcal{A} \leftarrow \mathcal{A} \cup \mathcal{A}_{\text{new}}, \quad \mathcal{R}_{\mathcal{A}} \leftarrow \mathcal{R}_{\mathcal{A}} \cup \mathcal{R}_{\text{new}}$
  Update `Tracker` with $\mathcal{A}$, $\mathcal{R}_{\mathcal{A}}$, and $\mathcal{Q}_{\text{train}}$
end while
Return: Final agent pool $\mathcal{A}$ and full response set $\mathcal{R}_{\mathcal{A}}$

Algorithm 2 Tracker Update Procedure

Input: Current agent pool $\mathcal{A}$, current agent responses $\mathcal{R}_{\mathcal{A}}$, and training questions $\mathcal{Q}_{\text{train}}$
for each question $i \in \mathcal{Q}_{\text{train}}$ **do**
  Compute question entropy $H_i(\mathcal{A})$
end for
for each mode **do**
  Compute variability score $V(\mathcal{A}_{\text{mode}})$
end for
Compute softmax probabilities from variability scores
Output: Variability scores, question entropies, and sampling probabilities.

B.4 Demonstration of P2P: An Example Run on Wave 42

To illustrate how P2P operates in practice, we present a detailed walkthrough of a single run on ATP Wave 42 (Trust in Science, 129 questions), tracing the system from the active endowment generation through final ensemble selection.

B.4.1 Experimental Setup

Questions are split according to a $7 : 1.5 : 1.5$ ratio into the training, validation, and test sets. We set the endowment budget to be 300, with 10 endowments generated for each mode during initial sampling. Active endowment generation is run in 10 update steps with the preset attribute bank, question patching (lowest 3; 75% endowment budget), and 10 attributes drawn for each endowment generation. Gemini-2.0-flash is used as the model backend.

B.4.2 Active Endowment Generation

Figure 5 displays the entropy change for each question across the update steps. Evidently, a significant number of questions experience a noticeable rise in entropy at the end of the first update step. From the second step onward, mixed mode enters the generation loop. Marginal gains in entropy are observed for the majority of questions, whereas some experience mild entropy drops. Figure 6 displays the question entropies at the end of generation.

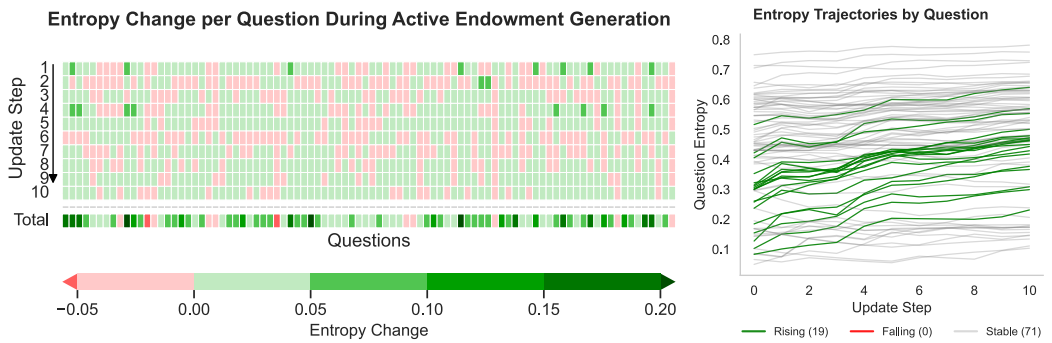


Figure 5: **Left:** Entropy change across update steps for each question during active endowment generation for ATP W42. Each column represents a question. Green indicates an increase in entropy from the previous step, while red indicates a decline. **Right:** Entropy trajectories by question (line chart). Active endowment generation progressively increases response diversity. Trends are classified based on slope and volatility of the entropy trajectory: "rising" if the slope exceeds 0.1 and standard deviation is above 0.02; "falling" if the slope is below –0.1 with sufficient volatility; otherwise labeled "stable".

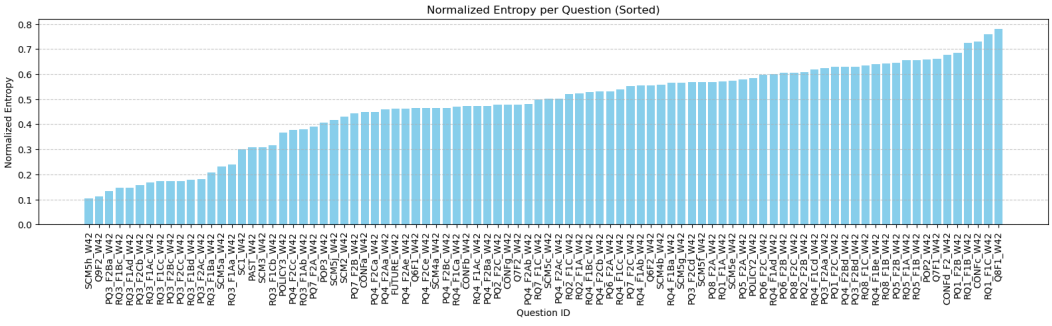


Figure 6: Question entropies at the end of active endowment generation.

B.4.3 Regression-based Aggregation

For regression analysis, categorical questions are binarized into question-option pairs. For the selection of the agent ensemble, we fit both a constrained lasso and a constrained elastic net, using CV to select the penalization parameters. Consequently, lasso selects 58 out of the 300 synthetic agents to form the agent ensemble, while elastic net selects 107. Figure 7 shows the hyperparameter selected for lasso using cross-validation.

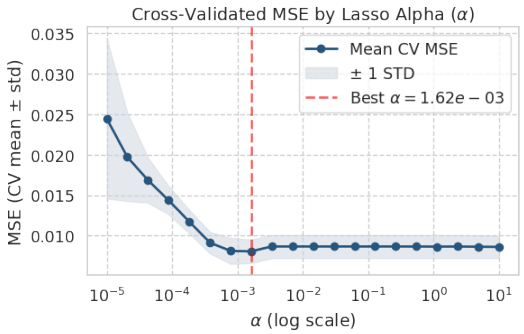


Figure 7: Cross-validated MSE by alpha (α), the penalization parameter, for constrained lasso. The best alpha selected by CV is 1.62×10^{-3}.

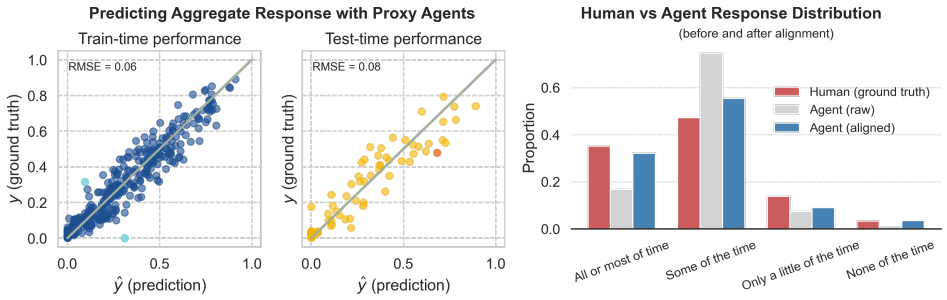


Figure 8: Prompts to Proxies: Emulation Results for ATP W42. **Left**: Train (left) and test (right) prediction performances for the binarized responses using constrained lasso. **Right**: Human and agent (pre- and post- alignment) aggregated responses for the sample test question `RQ4_F1Ae_W42`.

Figure 8 shows the prediction performances of the agent ensemble selected with constrained lasso. The train-time root mean squared error (RMSE) is 0.06, with test RMSE slightly higher at 0.08, indicating adequate generalization ability of the functional basis. The right panel of Figure 8 presents a snapshot for the aggregated responses to a test set question, comparing the human ground truth with agentic emulations before and after regression-based aggregation. Overall, the results are encouraging, especially considering that `P2P` constructs the endowments based solely on preset and learned attributes and is agnostic on the ground-truth demographic data. Figure 9 provides a breakdown of the agent ensemble selected by lasso. Notably, preset, freeform, and mixed templates all contribute non-trivially to the final ensemble. Figure 10 provides three examples of endowments generated and selected by `P2P` for the run on ATP W42, tracing back to their attributes, modes, and templates. Additionally, results of constrained elastic net are shown in Figures 11. Overall, the prediction results are similar to those using constrained lasso.

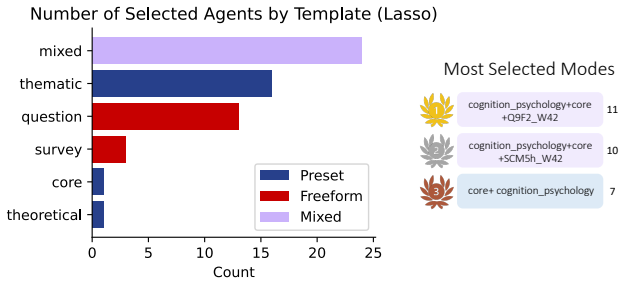


Figure 9: Breakdown of Selected Agents (58 in total) for W42 by Template. Preset, freeform, and mixed templates all contribute non-trivially to the final ensemble.

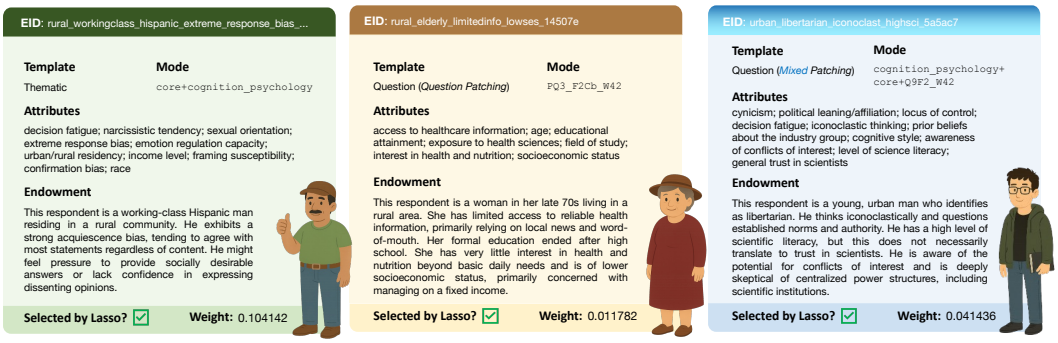


Figure 10: Examples of endowments generated by P2P for a run on ATP W42.

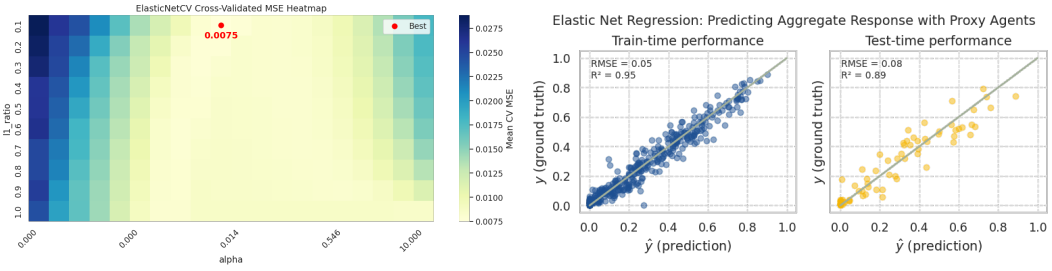


Figure 11: **Left:** Cross-validated MSE by alpha, the overall penalization parameter, and L1 ratio, the penalization weight for the L1 term, for constrained elastic net. The lowest cross-validated MSE is achieved at $\alpha = 6.95 \times 10^{-3}$, with l1 ratio set at 0.10. **Right:** Prediction performance of the agent ensemble (107 selected agents) using constrained elastic net. Results are similar to those using constrained lasso.

C Design Validation: Simulation Studies

In order to gauge the validity of the proposed preference model and the factors affecting alignment performance, we conduct three simulation studies where we have the privilege of ground-truth knowledge.

Simulation Design For the simulation studies, we use the American Trends Panel (ATP) Wave 42 (W42) as the survey source, covering the topic of public trust in science and scientists. We run the active endowment generation on the survey to get 300 persona endowments, and partition the survey questions into training, validation and test sets by the 0.7:0.15:0.15 ratio. We subsequently simulate the responses to the survey questions using endowed GPT-4o agents as the synthetic respondents. The ATP W42 dataset has in total 129 questions, which after binarization leads to 517 observations—85% of these observations form the train-and-validation (trainval) set and the rest 15% is used to test the predictive accuracy of the selected agent ensemble.

Construction of the Ground Truth Data To construct the ground truth data, we sample a fraction of the endowments as the ground truth and randomly assign weights to these endowments, which represent the proportions of the associated personas in the ground truth population. The ground truth data—percentages of individuals choosing difference options for a given question—are computed using weighted averages of responses provided by the ground-truth agents. The unselected endowments are used to construct the pool of proxy agents.

Regression Method For the simulation studies, we choose constrained lasso as our main regression alignment method for investigation. As a secondary choice, constrained elastic net is an extension of constrained lasso that trades off between the L1 and L2 penalty terms. It is expected to have more stable performance when the features (agent responses) are highly correlated. Our initial simulations indicate that constrained elastic net demonstrates near identical performance to constrained Lasso, partly because the latter is a special case of the former and we are using cross-validation to select the best penalization parameters. Therefore, for the sake of building intuitions, we focus on constrained lasso in the simulations and expect the results to be largely indicative of constrained elastic net's performance.

C.1 Recovering Ground Truth Agents with Constrained Lasso

As a first test of constrained lasso as a variable selection algorithm, we aim to evaluate its ability to recover ground-truth agents from a broader candidate pool. To render the selection task nontrivial, we set the prevalence of the ground-truth agents to 0.3—out of 10 selectable agents, only 3 contribute to the construction of the ground-truth signal. As the proxy agents dominate the agent pool, an adequate number of observations becomes essential for Lasso to reliably identify the true contributors. As a key metric, we define **observation-to-agent ratio (OAR)** to be the ratio of the number of training and validation observations to the number of selectable agents. By default, the OAR for our simulation baseline is 1.203.

To develop a panoramic view of constrained lasso's ability in ground truth recovery, we conduct multiple simulation rounds under varying OARs. We generate different OAR variations using two complementary subsampling strategies:

- **Subsampling endowments**. To increase the OAR, we adopt a subsampling strategy on endowments, where we subsample from the pool of selectable endowments, varying the total number of candidate agents while keeping the proportion of ground-truth agents fixed at 30%. The ground-truth weights are renormalized to sum to 1. We vary the subsample fraction from 0.5 to 1 in 20 equally spaced steps.
- **Subsampling questions**. To decrease the OAR, we subsample survey questions while keeping the original train-validation-test split and ratio unchanged. This mimics a setting with reduced behavioral signal for inference. As in the endowment-based approach, we use 20 subsample fractions ranging from 0.5 to 1.

We repeat each subsampling procedure 10 times to construct the final dataset used in our simulation study. Figure 12 shows the trends of the mean squared error of the test set and the precision—that is,

the proportion of selected agents who possess ground-truth endowments—across different OARs. To further illustrate the dynamics of selection, Figure 13 presents three representative snapshots of simulation runs, corresponding to low, medium, and high OAR settings.

As the figures indicate, constrained lasso's ability to recover ground-truth agents from the candidate pool improves with a higher observation-to-agent ratio. This is intuitive as more observations offer a richer signal of both individual-level idiosyncrasies and aggregate preferences. With abundant behavioral cues, the selection algorithm is better equipped to distinguish true contributors from correlated siblings. In contrast, when observations are sparse, Lasso tends to relinquish its sharpness as a selection tool and instead defaults to identifying a functional basis that best explains the limited data. The mild decline in predictive accuracy under low OAR regimes suggests that this functional basis still approximates the aggregate preference reasonably well, albeit imperfectly.

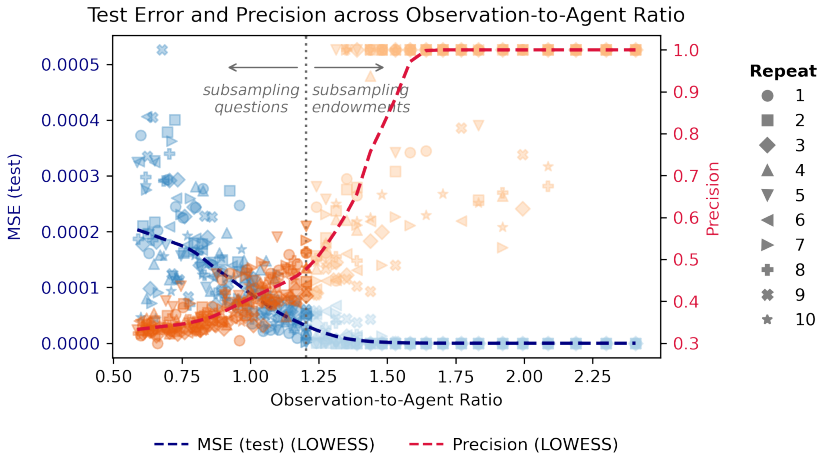


Figure 12: Mean Squared Error (Test Set) and Precision by Different Observation-to-Agent Ratio (OAR) for Constrained Lasso. Each simulation is repeated 10 times with different random seeds. Subsampling of endowments is done with fixed ground-truth-to-proxy ratio. Subsampling of responses is done with fixed train-valid-test split and ratio.

C.2 Emulating Ground Truth Responses with Proxy Agents

While the first simulation exercise has readily allured to constrained lasso's ability to find a functional basis to explain the observed data, in this second simulation study we extend the investigation by removing the ground-truth agents from the selectable agent pool. Under this setting, constrained lasso needs to fully rely on proxy agents to recover the ground-truth patterns. We analyze two distinct scenarios:

- **Subsampling proxies.** We fix the total number of observations and reduce the proxy pool by retaining only a fraction as selectable agents. We sweep across the fractions from 0.05 to 1 in 20 equal spaced steps.
- **Subsampling training and validation questions.** We fix the proxy pool but retain a fraction of training and validation observations. Similar to subsampling proxies, we construct 20 simulation rounds gradually raising the retained fraction from 0.05 to 1.

Like in Simulation Study 1, we repeat each simulation 10 times using different random seeds to form the final results. In Figure 14, the left panel displays the mean squared errors (MSE) and coefficients of determination (R^2) for the subsampling proxies scenario, while the right panel showcases the associated metrics for the subsampling training and validation questions scenario.

Notably, constrained lasso exhibits robust performance across all simulated conditions. Even with only 10 selectable proxy agents, it achieves a mean test MSE of 0.0067, yielding an R^2 of 0.93. Predictive accuracy improves as we increase the number of selectable proxy agents, plateauing at 0.99. In comparison, decreasing the number of observations has a lesser effect on predictive performance:

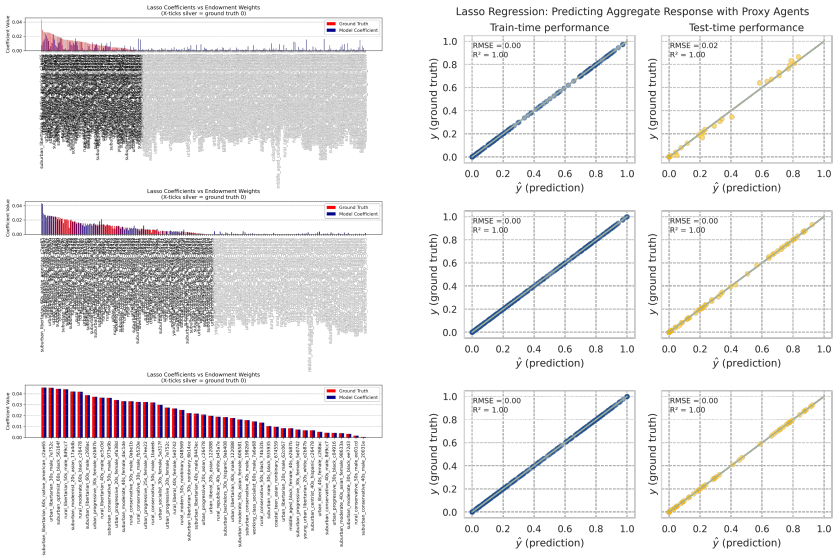


Figure 13: Snapshots of Weight-vs-Coefficient Comparison and Prediction Performance. Each row represents a simulation run. The red bars on the left panel represent ground truth agent weights used to generate the aggregate data and the blue bars Lasso coefficients. The observation-to-agent ratios for the top, middle and bottom panels are: 0.617, 1.203, 2.407. Lasso's ability to recover the ground truth agents from the agent pool degrades with lower observation-to-agent ratio, but its predictive accuracy only drops mildly.

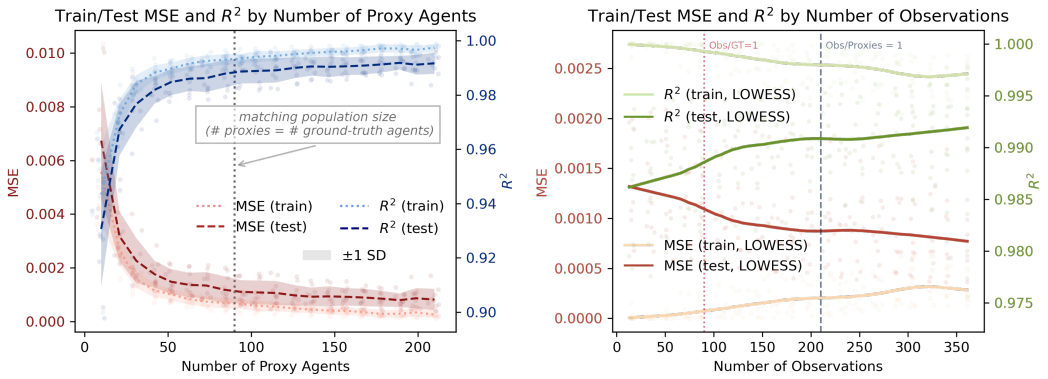


Figure 14: Left: Mean Squared Errors and Coefficients of Determination (R^2) by Number of Proxy Agents. Right: Mean Squared Errors and Coefficients of Determination (R^2) by Number of Observation. Each simulation is repeated 10 times with different random seeds.

in the recorded run with the lowest number of observations at 13, constrained lasso attains a test MSE of 0.0019 and an R^2 of 0.98. Figure 15 displays representative snapshots of the simulation runs.

The high predictive accuracy of the simulation runs indicates that constrained lasso can efficiently find a functional basis from a select few agents and determine the agent weights to emulate the ground-truth data patterns based on limited observations. This is in fact not surprising because in the initial assignment of the endowments into the ground truth and proxy groups we have used a random partition. As informed by our proposed behavioral preference model, this causes the ground truth vectors—used in the simulation study to form the aggregate preference—and proxy vectors to span the same preference space. As the preference space is not directly accessible, we can derive a partial gauge through the observed agent responses. If the responses of a group of agents have higher variability, it is an indication that the preference space they span is large. Conversely, if agents tend

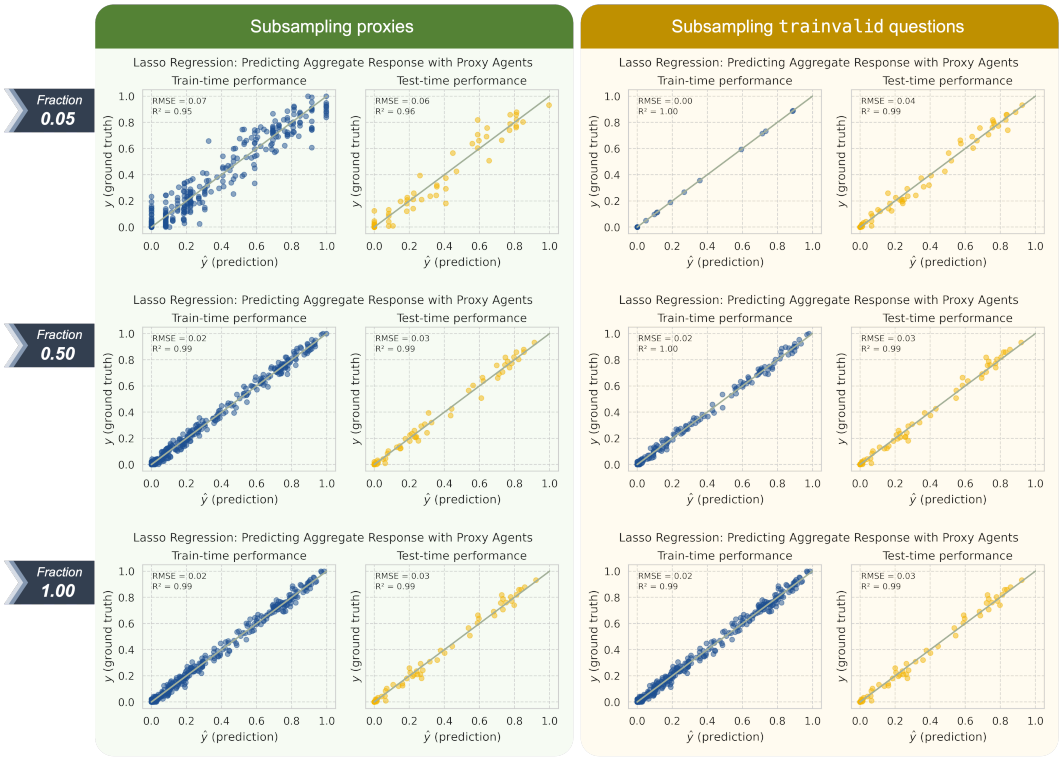


Figure 15: Snapshots of Prediction Performance for Simulation Study 2.

to agree on their responses to different questions, it is a signal that their associated vectors in the preference space might overlap in sub-dimensions, leading to a pool generalization ability.

To effectively measure response variability, in the paper we introduce the notion of **question entropy**:

$$H_i(\mathcal{A}) = \frac{-\sum_{k=1}^{K_i} p_{ik} \log_2 p_{ik}}{\log_2 K_i} \tag{9}$$

where $H_i(\mathcal{A})$ is the normalized entropy of question i based on responses from a group of agents $\mathcal{A}$, K_i denotes the number of unique response options $\{1, \ldots, K_i\}$ for question i, and p_{ik} represents the empirical proportion of responses selecting option k. The normalization factor ensures comparison across questions. A high question entropy indicates that the responses are highly varied for this question.

For the simulation studies, we define **group entropy** as the average question entropy computed using responses from that group:

$$H(\mathcal{A}) = \frac{1}{N} \sum_{i=1}^{N} H_i(\mathcal{A}). \tag{10}$$

The group entropy is calculated using responses to all survey questions, regardless of the split. Based on it, we introduce the notion of **Entropy Coverage Ratio (ECR)** defined as the ratio between the group entropy of proxy agents and that of the ground-truth agents:

$$ECR = \frac{H(\mathcal{A}_{\text{proxy}})}{H(\mathcal{A}_{\text{gt}})}. \tag{11}$$

An ECR lower than 1 indicates that the proxy agents have less response variability than the ground truth, while an ECR greater than 1 indicates the opposite.

The left panel of Figure 16 shows the distribution of entropy coverage ratios (ECRs) for the two analyzed scenarios. The mass is concentrated around 1.0, corroborating the earlier observation that random partitioning tends to preserve entropy structure. The right panel plots ECR against

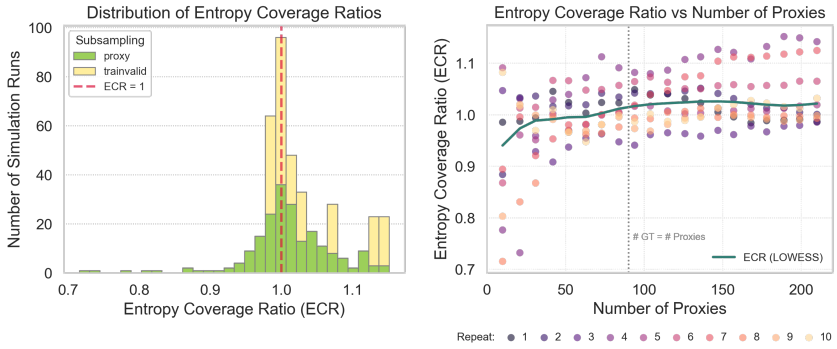


Figure 16: Entropy Coverage Ratio Diagnostics for Simulation Study 2.

the number of proxies for the proxy-subsampling scenario. The relationship between ECR and the number of proxies is subtly positive: when the number of proxies is small, the likelihood of drawing an unrepresentative sample increases. As the sample size grows, group-level entropy converges toward the population-level entropy—an empirical manifestation of the law of large numbers.

C.3 Response Variability Matters

In this final simulation study, we analyze the effect of the entropy coverage ratio on predictive accuracy for constrained lasso.

Figure 17 displays the group entropies for THE W42 endowments organized by modes—we call them **mode entropies** in the endowment generation logic. We subsume them further into three different tiers according to the group entropy value: in total, we have 71 endowments in the low entropy tier, 92 in the mid entropy tier, and 137 in the high entropy tier.

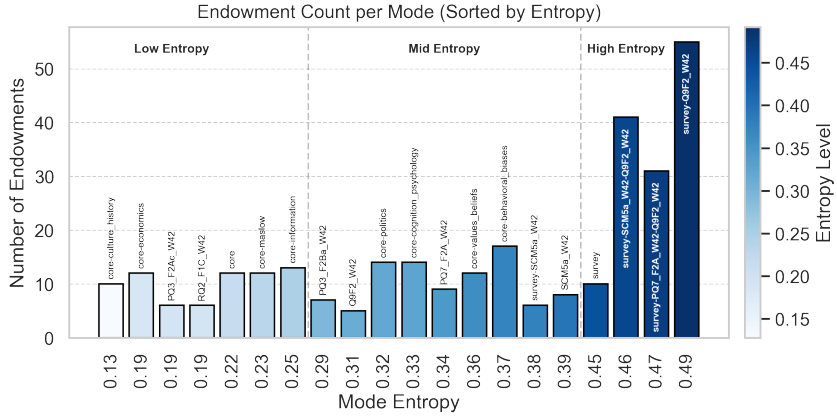


Figure 17: Endowment Count per Mode (Sorted by Entropy).

To cover a wide range of ECR ratios, we design the following simulation strategy:

1. **Ground Truth Construction.** For each simulation round, pick one of the three tiers as the ground-truth tier and sample 30 endowments from the tier to construct the ground-truth agents and assign ground-truth weights. The unselected endowments from the tier are joined with the endowments from the other two tiers to form the pool of proxy candidates.
2. **Proxy Agents Construction.** For each simulation run, we select 30 proxy agents to form the proxy pool. We begin by selecting agents from the lowest entropy modes (ensuring that the cumulative candidate endowments surpass 30), and in each consecutive run, include the next mode with higher entropy into the selectable modes. We sweep across the modes by increasing mode entropy.

3. **Proxy Agents Selection and Response Emulation.** In each simulation run, after the proxy pool has been constructed, we use constrained lasso to select proxy agents to emulate the observed aggregate response data.

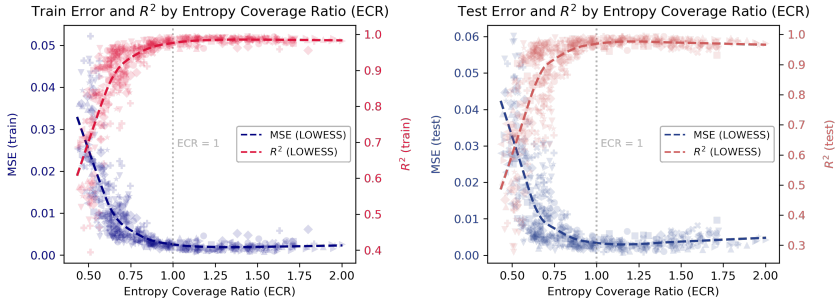


Figure 18: Mean Squared Error and Coefficient of Determination (R^2) by Entropy Coverage Ratio (ECR) for the trainval set (left) and test test (right). Entropy Coverage Ratio is defined as the ratio of the average response entropies of the proxies and ground-truth agents. LOWESS curves are fit using 10 repetitions.

We repeat each simulation 10 times to form the final results. Figure 18 visualizes the train and test performance of the constrained lasso by the entropy coverage ratio. As the ECR increases, the predictive accuracy improves at both the training and the test times, lending further credence to the analysis based on the preference model in the second simulation study.

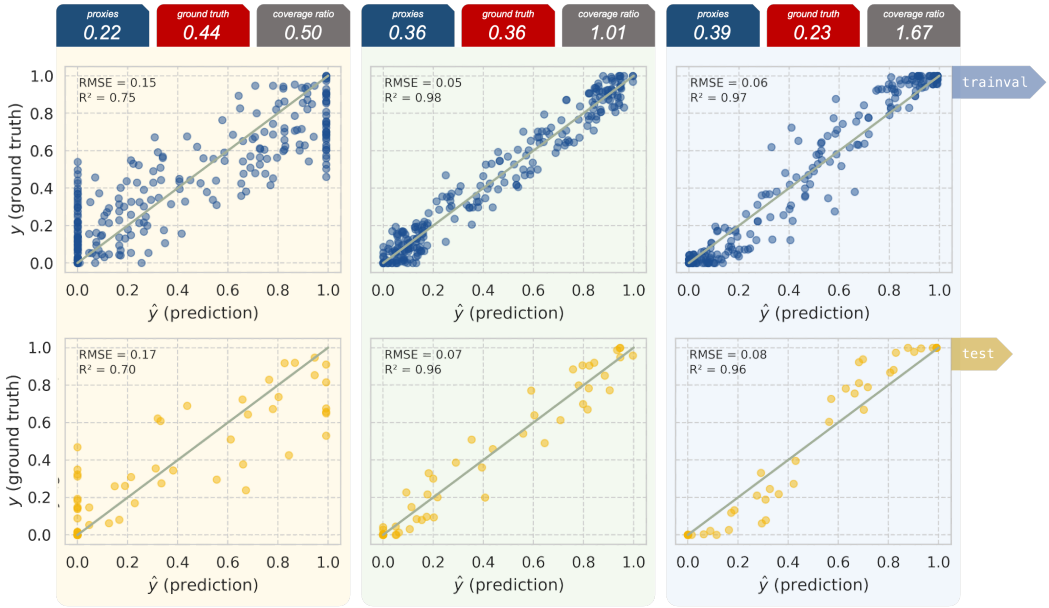


Figure 19: Snapshots of Prediction Performance for Simulation Study 3.

Figure 19 offers three snapshots of the simulation runs corresponding to ECRs equaling 0.50, 1.01 and 1.67. When proxy entropy is low compared with ground truth, the proxy agents fail to provide a sufficiently expressive basis to span the aggregate ground-truth preference. As a result, constrained Lasso can only approximate a lower-dimensional projection of the true preference signal. This results in poor prediction performance, as illustrated in the leftmost snapshot. Conversely, when proxy entropy exceeds ground-truth entropy, prediction error begins to increase—albeit at an infinitesimal scale. The subtle deterioration of performance when ECR surpasses 1 could be due to various reasons. A plausible cause is from the data perspective: when the proxy pool becomes too diverse, it may introduce spurious variability that does not align well with the true underlying structure. Put differently, an observed response might amalgamate signals from different latent preference factors,

some of which are irrelevant to the aggregate preference in focus. While greater diversity expands the representational capacity of the proxy pool, it may also dilute the signal and induce overfitting, thereby reducing generalization accuracy. In practice, we posit that over-diversity is less likely than under-diversity, as existing studies using demographic conditioning [4, 47] indicate LLM agents do not match closely with human ground truth. This justifies optimizing response variability in Stage 1 of P2P.

The key takeaway from the simulation study is that absent the knowledge of ground truth, it is crucial to ensure an adequate group entropy among the proxy agents to form a functional basis for ground-truth emulation. However, one should also refrain from solely relying on entropy as a measure of preference diversity lest spurious correlations beguile us under limited data regimes. While our alignment method is intentionally designed to function with minimal data—requiring only aggregate response data—in practice, if individual human responses are available, a practitioner may use the group entropy computed from these responses as an anchor for endowment generation.

D Empirical Experiments: Supplementary Results

D.1 Main Results: Baseline Comparison on ATP Waves

In our setup, MSE is computed on the option-level probabilities after binarizing each multiple-choice item, so MSE penalizes misallocation of probability mass across the full categorical distribution. To facilitate comparisons across works and corroborate the use of MSE as a valid distribution metric, we follow previous studies [8, 20, 17] and additionally report Jensen-Shannon Divergence (1-JSD) and Earth Mover Distance (EMD) for the key results in this paper.

Table 3 presents the full distribution metrics comparison of baselines and P2P across 14 ATP waves. P2P consistently outperforms Vanilla and PERSONA across all three metrics, registering lower MSE (0.014 vs. 0.025–0.029), higher 1-JSD (0.830 vs. 0.759–0.771), and lower EMD (0.066 vs. 0.085–0.091). The 23–27% reduction in EMD indicates that P2P predictions are structurally closer to ground-truth distributions.

Table 3: Full distribution metrics comparison across 14 ATP waves. MSE measures mean squared error of predicted response probabilities. 1-JSD (1 minus Jensen-Shannon Divergence) measures distributional similarity (higher is better). EMD (Earth Mover's Distance) measures distributional distance (lower is better). Each cell reports the mean over three runs, with the sample SD in parentheses below.

	Vanilla			PERSONA			P2P		
Wave	MSE	1-JSD	EMD	MSE	1-JSD	EMD	MSE	1-JSD	EMD
W26	.015	.820	.072	.024	.772	.082	.010	.868	.056
	(.002)	(.011)	(.004)	(.003)	(.009)	(.004)	(.000)	(.001)	(.001)
W27	.032	.795	.086	.041	.728	.104	.022	.823	.071
	(.004)	(.008)	(.007)	(.001)	(.004)	(.001)	(.001)	(.006)	(.003)
W29	.026	.767	.095	.030	.748	.106	.016	.815	.083
	(.003)	(.009)	(.002)	(.002)	(.008)	(.005)	(.005)	(.026)	(.013)
W32	.019	.804	.067	.025	.772	.082	.016	.832	.065
	(.000)	(.006)	(.001)	(.000)	(.004)	(.001)	(.002)	(.007)	(.003)
W34	.021	.777	.084	.018	.790	.080	.011	.822	.070
	(.004)	(.015)	(.009)	(.001)	(.010)	(.006)	(.001)	(.010)	(.007)
W36	.033	.747	.085	.028	.787	.074	.019	.821	.061
	(.001)	(.011)	(.005)	(.007)	(.011)	(.002)	(.004)	(.016)	(.008)
W41	.022	.761	.101	.027	.756	.102	.013	.828	.069
	(.001)	(.003)	(.003)	(.003)	(.018)	(.009)	(.001)	(.002)	(.006)
W42	.029	.767	.081	.036	.753	.090	.011	.853	.053
	(.005)	(.010)	(.004)	(.001)	(.003)	(.001)	(.000)	(.003)	(.001)
W45	.015	.805	.072	.029	.773	.085	.008	.848	.054
	(.003)	(.011)	(.007)	(.001)	(.006)	(.001)	(.002)	(.013)	(.010)
W49	.046	.734	.093	.041	.746	.089	.021	.814	.065
	(.001)	(.013)	(.004)	(.006)	(.013)	(.001)	(.004)	(.017)	(.004)
W50	.027	.710	.100	.026	.699	.112	.013	.805	.075
	(.003)	(.019)	(.006)	(.002)	(.011)	(.001)	(.004)	(.015)	(.008)
W54	.020	.753	.085	.024	.748	.080	.014	.801	.068
	(.003)	(.021)	(.009)	(.003)	(.009)	(.002)	(.002)	(.019)	(.007)
W82	.021	.761	.088	.021	.772	.081	.008	.855	.052
	(.004)	(.016)	(.007)	(.002)	(.006)	(.002)	(.002)	(.012)	(.005)
W92	.026	.800	.089	.031	.780	.101	.017	.834	.076
	(.002)	(.006)	(.004)	(.001)	(.003)	(.002)	(.003)	(.009)	(.005)
Avg	.025	.771	.085	.029	.759	.091	**.014**	**.830**	**.066**

Across the three runs, sample SDs are small relative to the differences among the methods: at most .007 for MSE, .026 for 1-JSD, and .013 for EMD (Table 3). For the entropy and cost in Table 1, entropy SDs are at most .015 (Vanilla), .007 (PERSONA), and .025 (P2P), and cost SDs at most

$0.007 for both baselines and $0.26 for P2P. The latter comes from W41, where one run generated 496 instead of 300 agents after repeated connection retries during generation.

Table 4 compares lasso and elastic net as Stage 2 regression choices for P2P. Performance is nearly identical across all waves (average Δ MSE ≈ 0.0001, with 9 of 14 waves showing no difference). As lasso typically selects a more compact ensemble than elastic net, we prioritize its use as the default regression method.

Runtime Runtime is dominated by API inference. By design, P2P parallelizes both prompt generation and survey response elicitation, but the exact runtime varies by `update_steps`, `max_workers`, API rate limits and connection stability. For reference, a typical run on a single ATP wave (20 workers, 5 update steps, Gemini-2.0-Flash) takes about 26 mins, with regression under 2 minutes.

Table 4: Comparison of lasso and elastic net regression for P2P across 14 ATP waves. Values are mean $\pm$ sample SD over three runs. Both methods yield nearly identical performance, demonstrating robustness to regularization choice.

Wave	Lasso MSE	ENet MSE	Δ	Winner
W26	.0098 $\pm$ .0002	.0098 $\pm$ .0002	.0000	Tie
W27	.0218 $\pm$ .0010	.0238 $\pm$ .0027	.0020	Lasso
W29	.0156 $\pm$ .0049	.0156 $\pm$ .0049	.0000	Tie
W32	.0160 $\pm$ .0019	.0160 $\pm$ .0019	.0000	Tie
W34	.0111 $\pm$ .0011	.0111 $\pm$ .0011	.0000	Tie
W36	.0191 $\pm$ .0042	.0187 $\pm$ .0037	.0004	ENet
W41	.0126 $\pm$ .0009	.0126 $\pm$ .0009	.0000	Tie
W42	.0110 $\pm$ .0003	.0101 $\pm$ .0006	.0009	ENet
W45	.0079 $\pm$ .0023	.0079 $\pm$ .0023	.0000	Tie
W49	.0211 $\pm$ .0037	.0211 $\pm$ .0037	.0000	Tie
W50	.0133 $\pm$ .0039	.0133 $\pm$ .0039	.0000	Tie
W54	.0144 $\pm$ .0018	.0146 $\pm$ .0018	.0002	Lasso
W82	.0083 $\pm$ .0016	.0083 $\pm$ .0016	.0000	Tie
W92	.0168 $\pm$ .0029	.0173 $\pm$ .0023	.0005	Lasso
Avg	**.0142**	**.0143**	.0001	–

D.1.1 Clustering

Based on Vanilla and PERSONA, we design two variants using K-means clustering as a pre-processing step. Specifically, we first expand the initial pool of personas for Vanilla and PERSONA to 1000, computing their embeddings (via OpenAI's text-embedding-3-large) and cluster them (k=300). We pick the personas (medoids) closest to the centroids to construct the candidate pool and then run our regression stage on the resulting pool. Table 5 shows results on ATP W42.

Table 5: Comparison of Vanilla, PERSONA and their clustering-based variants where 1000 agents are reduced to 300 via k-means clustering.

Metric	**Vanilla**	**Vanilla (1000→300)**	**PERSONA**	**PERSONA (1000→300)**
Avg. Entropy ↑	0.3172 $\pm$ 0.0069	0.3161	0.2871 $\pm$ 0.00026	0.2844
Test MSE (Lasso) ↓	0.0285 $\pm$ 0.00490	0.0266	0.0362 $\pm$ 0.00116	0.0395
Test MSE (Enet) ↓	0.0286 $\pm$ 0.00476	0.0266	0.0339 $\pm$ 0.00124	0.0387
Cost (USD) ↓	**0.8649 $\pm$ 0.0070**	0.8833[†]	1.7129 $\pm$ 0.000016[*]	1.7203[*†]

[*] Only considering survey elicitation cost on 300 personas—actual cost is higher.
[†] Excluding embedding cost.

Clustering yields a test MSE of 0.0266 for Vanilla that lies well within the variability band of the original Vanilla baseline (0.0285 $\pm$ 0.0049) and slightly worse MSE for PERSONA (0.0395 vs. 0.0362 $\pm$ 0.0011), with no conspicuous change on entropy and a slight increase in cost (for Vanilla due to extra generation). Hence, clustering does not yield a stronger baseline.

This matches our intuition for two reasons. First, when we apply clustering on top of the existing baselines (starting from the same 300 endowments/personas), it can only discard information that the regression step could have used. In other words, it weakens the candidate pool without giving L1 regression any new capability. Second, even in the more lenient setting where we give clustering a head start by generating 1000 candidates and then compressing them to 300, the mechanism still does not produce a stronger baseline. As clustering operates in the embedding space and tends to group personas by textual or stylistic similarity, it does not necessarily track preference-relevant variation in survey responses. As a result, minority or low-density preference types are prone to being averaged away in embedding clustering, whereas our regression layer selects proxies directly based on their contribution to matching the observed response distributions. This corroborates the design of our main baselines.

D.1.2 Raw Model Performance

To better contextualize the gain of the ensemble methods (Vanilla, PERSONA, and P2P), we provide the raw model backend performance as a reference baseline. To construct the baseline, we directly query the model backend (gemini-2.0-flash) for answers to the survey questions, using the following instruction:

```
You are completing a public opinion survey.  Answer each
question as an average adult respondent in the US population,
based only on the question text and options.  Respond with
only the final answer string, not the code or label in
brackets.  Do not include any reasoning, explanation, or
commentary.  Do not preface your answer with phrases like
'I would choose'.  Just return the answer text exactly as it
appears in the options.
```

We draw 300 Monte-Carlo samples per question with temperature set to 0.7. This results in an empirical answer distribution per question. On ATP W42, this baseline attains test MSE 0.0918 and average entropy 0.0161, conspicuously worse than the results for the ensemble baselines, confirming the gains via ensemble methods (Vanilla, PERSONA, and P2P).

D.2 Locale Generalization: World Values Survey

The main paper results focus on the American Trends Panel (ATP), which represents a single survey program targeting the US population. While our 14-wave study demonstrates that P2P generalizes across multiple topics and different cross sections, it remains intriguing how it performs at other locales and on different survey programs. To address this question, we evaluate P2P on the World Values Survey (WVS) Wave 7 [23], using cleaned data from Cao et al. [8]. We select three locales representing different cultural and linguistic contexts: **United States**, as it provides a bridge to our ATP findings, enabling direct comparison within the same population under a different survey program; **Great Britain**, which represents Western Europe, sharing linguistic and value overlap with the US but differing in political and cultural contexts; and **Hong Kong SAR**, a multicultural hub in East Asia, allowing us to test on populations less represented in typical LLM training data.

D.2.1 Experiment Setup

For the WVS experiment, we use a similar setup to the ATP study: survey questions are split randomly according to a $7 : 1.5 : 1.5$ ratio into the training, validation, and test sets; endowment budget is set at 300; initial sampling generates 10 endowments for each mode; active endowment generation is run in 5 update steps with the preset attribute bank, question patching (lowest 3; 75% endowment budget); 10 attributes drawn for each endowment generation; Gemini-2.0-flash is used as the backend. Different from the ATP study, we additionally pass the locale information (US, GB, HK) to the endowment model to ensure regional relevance of generated personas.

Notably, our random split protocol differs from Cao et al. [8]'s Q1+Q2→Q3 topic shift protocol (Appendix D.4). This is intentional, as it isolates the effect of locale from confounding topic shift effects, providing a cleaner assessment of cross-locale generalizability.

Table 6 reports results across the three locales. Each experiment is repeated 3 times.

D.2.2 Results and Discussion

Table 6: P2P performance on WVS Wave 7 (US, Great Britain, Hong Kong SAR) under a 7:1.5:1.5 question split, using gemini-2.0-flash as the backend. Reported are mean ± sample SD over three repeated runs.

Metric	WVS W7 US	WVS W7 GB	WVS W7 HK
Avg. Entropy ↑	0.5720 ± 0.0149	0.5710 ± 0.0044	0.5478 ± 0.0211
Cost (USD) ↓	1.7551 ± 0.0097	1.7095 ± 0.0031	1.7365 ± 0.0080
Test MSE (Lasso) ↓	0.0149 ± 0.0005	0.0141 ± 0.0014	0.0242 ± 0.0013
Test 1-JSD (Lasso) ↑	0.8021 ± 0.0042	0.7892 ± 0.0052	0.7532 ± 0.0056
Test MCPA (Lasso) ↑	0.6667 ± 0.0152	0.6481 ± 0.0321	0.5789 ± 0.0263
Test EMD (Lasso) ↓	0.1073 ± 0.0032	0.1032 ± 0.0072	0.1423 ± 0.0035
Test MSE (Elastic Net) ↓	0.0147 ± 0.0005	0.0137 ± 0.0010	0.0244 ± 0.0005
Test 1-JSD (Elastic Net) ↑	0.8056 ± 0.0023	0.7950 ± 0.0073	0.7555 ± 0.0051
Test MCPA (Elastic Net) ↑	0.6930 ± 0.0152	0.6296 ± 0.0321	0.5877 ± 0.0304
Test EMD (Elastic Net) ↓	0.1064 ± 0.0022	0.1004 ± 0.0054	0.1431 ± 0.0044

Table 7: Comparison of P2P performance on ATP (US) and WVS (US). ATP values are mean ± sample SD over 14 waves; WVS values are mean ± sample SD over 3 runs.

Metric	ATP (14 waves)	WVS US
Test MSE (Lasso) ↓	0.014 ± 0.004	0.015 ± 0.001
Avg. Entropy ↑	0.53 ± 0.06	0.57 ± 0.01
Cost (USD) ↓	0.8 ± 0.2	1.76 ± 0.01

As shown in Table 6, performance is higher for US (MSE ≈ 0.015) and GB (MSE ≈ 0.014) and somewhat lower for HK (MSE ≈ 0.024), aligning with broader findings that large language models trained primarily on English and Western-centric data tend to better match Western public opinion than non-Western populations [17]. The HK performance gap motivates the error analysis presented in Section 4.4 of the main paper. Despite the gap, the HK results lie just above the performance range (0.008–0.022) observed across our 14-wave ATP study presented in the main paper (cf. Table 1), suggesting that while locale affects difficulty, P2P remains broadly applicable. Compared with ATP, WVS results register a higher per-survey cost at approximately $1.73, due to the higher number of survey questions (244 for US and HK, 232 for GB). Notably, the US results are nearly identical across survey programs (ATP: 0.014; WVS: 0.015), providing evidence against program-specific overfitting (Table 7).

D.3 Ablation Studies

D.3.1 Attributes

As discussed in Section 3.2, attributes play the role of control handles in P2P's generation logic, and *freeform* templates allow P2P to derive attributes from a specific survey or question, complementing the preset attribute bank with data-driven insights on endowment generation. To evaluate the potency of the freeform templates, we conduct an ablation in study in which we sequentially turn off question patching, survey patching and both. We additionally test the impact of per-endowment attribute cap through a setting where the default max attribute=10 is changed to 20.

Results in Table 8 suggest that performance of P2P is robust to the attribute cap: entropy is essentially unchanged, and the difference in test MSE falls within run-to-run variability. Turning off freeform templates (especially question patching) degrades both average entropy and test MSE without meaningful cost gain. Therefore, from a performance vantage, it is recommended to keep the freeform templates for the extra benefits of data-driven insights.

Table 8: Attribute ablation. Reported are mean ± sample SD over three repeated runs.

Metric	**Avg. Entropy ↑**	**Test MSE ↓ (Lasso)**	**Test MSE ↓ (Enet)**	**Cost (USD) ↓**
Original	**0.4451** ± .0256	**0.0087** ± .0026	**0.0082** ± .0025	0.9708 ± .0036
Survey patch off	0.4378 ± .0134	0.0104 ± .0024	0.0102 ± .0017	0.9785 ± .0122
Question patch off	0.3989 ± .0203	0.0114 ± .0016	0.0109 ± .0015	0.9698 ± .0011
Freeform off	0.3953 ± .0097	0.0128 ± .0008	0.0123 ± .0012	**0.9688** ± .0015
max attributes = 20	0.4363 ± .0061	0.0124 ± .0046	0.0115 ± .0041	0.9880 ± .0167

D.3.2 Endowment Budget

We run P2P for 9 different endowment budgets, scaled linearly from 130 to 450. We increase the update steps of active endowment generation accordingly to ensure that at each step the number of new endowments generated is fixed at 20. We also include a budget of 110 with no updates as an ablation of adaptive sampling. Table 9 shows that on W42, a higher budget generally improves test performance but gains plateau after about three updates, indicating a moderate budget suffices.

Table 9: Performance of P2P with varying endowment budget. The average question entropy, test MSE (Lasso and ElasticNet), and generation cost are reported. All values are mean ± sample SD over three repeated runs.

(a) Budgets 110–210

Endowment Budget	110	130	170	210
Avg. Entropy ↑	0.3777 ± 0.0191	0.3697 ± 0.0057	0.4180 ± 0.0080	0.4341 ± 0.0043
Test MSE (Lasso) ↓	0.0144 ± 0.0018	0.0167 ± 0.0041	0.0100 ± 0.0025	0.0111 ± 0.0006
Test MSE (Enet) ↓	0.0149 ± 0.0014	0.0156 ± 0.0028	0.0103 ± 0.0026	0.0104 ± 0.0006
Cost (USD) ↓	**0.3506** ± 0.0033	0.4154 ± 0.0054	0.5484 ± 0.0058	0.6752 ± 0.0098

(b) Budgets 250–330

Endowment Budget	250	290	330
Avg. Entropy ↑	0.4179 ± 0.0165	0.4369 ± 0.0168	0.4308 ± 0.0208
Test MSE (Lasso) ↓	0.0124 ± 0.0043	0.0090 ± 0.0013	0.0133 ± 0.0030
Test MSE (Enet) ↓	0.0123 ± 0.0045	0.0097 ± 0.0017	0.0121 ± 0.0031
Cost (USD) ↓	0.8043 ± 0.0032	0.9365 ± 0.0027	1.0688 ± 0.0119

(c) Budgets 370–450

Endowment Budget	370	410	450
Avg. Entropy ↑	0.4216 ± 0.0179	**0.4540** ± 0.0142	0.4433 ± 0.0092
Test MSE (Lasso) ↓	0.0140 ± 0.0037	**0.0080** ± 0.0006	0.0099 ± 0.0022
Test MSE (Enet) ↓	0.0136 ± 0.0055	**0.0084** ± 0.0004	0.0096 ± 0.0019
Cost (USD) ↓	1.2050 ± 0.0069	1.3295 ± 0.0034	1.4619 ± 0.0137

D.3.3 Model Backend

We test P2P's performance on six model backends: GPT-4.1-mini, GPT-4.1-nano, Gemini-2.0-flash, Gemini-2.5-flash-lite, DeepSeek-V4-Flash and a locally hosted Qwen. Performance varies substantially, with the best average question entropy (0.471) achieved by GPT-4.1-mini (Table 10). Gemini-2.0-flash achieves the lowest test MSE (0.0087 with lasso) at about one quarter of the cost of GPT-4.1-mini. DeepSeek-V4-Flash is the cheapest backend and ranks second on entropy, but with an intermediate test MSE (0.0165 with lasso). Smaller models, like GPT-4.1-nano and Gemini-2.5-flash-lite, perform worse, consistent with our theoretical intuition that limited model capacity constrains the learnability condition (Definition 2.2).

Table 10: Comparison of P2P performance on W42 using different model backends. Reported are average question entropy, test MSE (Lasso and ElasticNet), and generation cost. For each model, we fix the endowment budget at 300. Values are means over three repeated runs, with sample SDs in parentheses.

Model	Avg. Entropy	Test MSE (Lasso)	Test MSE (Enet)	Cost (USD)
GPT-4.1-mini	**0.4708**	0.0090	0.0088	3.8342
	(0.0168)	(0.0014)	(0.0014)	(0.0449)
GPT-4.1-nano	0.3996	0.0226	0.0213	0.9265
	(0.0154)	(0.0042)	(0.0048)	(0.0043)
Qwen	0.1932	0.0209	0.0201	–
	(0.0403)	(0.0035)	(0.0029)	
Gemini-2.0-flash	0.4451	**0.0087**	**0.0082**	0.9708
	(0.0256)	(0.0026)	(0.0025)	(0.0036)
Gemini-2.5-flash-lite	0.3608	0.0149	0.0151	1.0225
	(0.0084)	(0.0009)	(0.0013)	(0.0100)
DeepSeek-V4-Flash	0.4463	0.0165	0.0157	**0.7234**
	(0.0073)	(0.0035)	(0.0051)	(0.0153)

Note. Qwen was run on a local server, thereby incurring no API cost.

D.3.4 Regression-based aggregation

To gauge the contribution of the regression-based aggregation module, we ablate it by replacing regression with simple averaging of agent responses, which increases the test MSE from 0.0087 to 0.0332 (Table 11). In essence, regression not only distills the agent collection into a more compact ensemble, but it also plays a crucial role in ensuring the population preference reflected in the survey data are actually learned by the ensemble.

Table 11: Comparison of P2P performance with and without the regression-based aggregation stage on ATP W42. In the ablation, agent responses from active endowment generation are combined by simple averaging, while the regression setting uses lasso. Reported are mean ± sample SD over three repeated runs.

Metric	Without regression	With regression
Test MSE ↓	0.0332 ± 0.0024	**0.0087** ± 0.0026

We also compare the default option-level weighting against block weighting, where each question contributes equally to the loss. Performance remains comparable under both schemes for W42 (Table 12).

Table 12: Sensitivity check of weighting schemes. Reported are mean ± sample SD over 3 runs.

Weighting Scheme	**Train MSE ↓ (Lasso)**	**Test MSE ↓ (Lasso)**	**Train MSE ↓ (Enet)**	**Test MSE ↓ (Enet)**
Default	0.00497 ± .00156	0.00870 ± .00256	**0.00403 ± .00144**	0.00821 ± .00249
Block	**0.00475 ± .00232**	**0.00868 ± .00257**	0.00442 ± .00236	**0.00806 ± .00278**

D.3.5 Stage 2 Configuration

P2P's Stage 2 uses regression algorithms to select the agent ensemble. Users have considerable latitude in configuring the regression algorithm, including, inter alia, the choice of constraints and the inclusion of additional anchor features. Before computing the distributional metrics (1-JSD, EMD, MCPA), renormalization is performed by design to ensure that the final answers constitute valid probability distributions. To illustrate the implications of these design choices, we present

an analysis and ablation study examining the effects of the mass penalty constraints and additional anchor features.

Mass penalty and renormalization Recall that j indexes agents. Write $\hat{r}_{ik}(\mathbf{w}) = \sum_{j=1}^{N_A} w_j d_{j,ik}$ for the predicted share of option k of question i, and $m_i(\mathbf{w}) = \sum_{k=1}^{K_i} \hat{r}_{ik}(\mathbf{w})$ for the predicted *mass* of question i. The constraint in Eq. (5) requires every question's mass to equal one:

$$\min_{\mathbf{w}} \; L(\mathbf{w}) + \alpha\|\mathbf{w}\|_1 \quad \text{s.t.} \quad m_i(\mathbf{w}) = 1 \quad \forall i \in \mathcal{Q}_{\text{train}}, \tag{12}$$

where $L(\mathbf{w}) = \frac{1}{2n}\sum_i \sum_k (r_{ik} - \hat{r}_{ik}(\mathbf{w}))^2$ is the squared loss over the n training option rows. In most cases this is equivalent to a single sum-to-one constraint on the weights $\mathbf{w}$. When every agent answers every question exactly once, $\sum_k d_{j,ik} = 1$ for all i and j, so $m_i(\mathbf{w}) = \sum_j w_j$ for every question and the $|\mathcal{Q}_{\text{train}}|$ constraints reduce to

$$\min_{\mathbf{w}} \; L(\mathbf{w}) + \alpha\|\mathbf{w}\|_1 \quad \text{s.t.} \quad \sum_{j=1}^{N_A} w_j = 1. \tag{13}$$

The two constraints differ in their treatment of edge cases: e.g., (i) when an agent leaves a question unanswered while remaining in the basis, so that its answers no longer sum to one within that question, and (ii) when the fit includes an intercept (below), which contributes an option-count-dependent term to each question's mass. The default pipeline of `P2P` adopts column-drop cleaning, which removes an agent with a missing or invalid answer from the candidate pool before the selection step, thereby ruling out case (i). Other cleaning schemes, for example ones that retain partially answering agents, would fall into case (i). By imposing a constraint per question, (12) handles these edge cases more flexibly than (13), but simultaneously it engenders a larger set of equality constraints that is difficult to optimize. Rather than imposing (12) as hard constraints, `P2P` relaxes it into a quadratic *mass penalty*:

$$\min_{\mathbf{w}} \; L(\mathbf{w}) + \alpha\|\mathbf{w}\|_1 + \lambda \sum_{i \in \mathcal{Q}_{\text{train}}} \big(m_i(\mathbf{w}) - 1\big)^2, \qquad \lambda = 100. \tag{14}$$

On the main panel (Table 1), it keeps m_i between 0.911 and 1.000. Therefore, predictions are approximately but not exactly valid distributions. For the distributional metrics (1-JSD, EMD, MCPA), each question's predicted shares are clipped at zero and renormalized to sum to one before scoring, so validity does not depend on the implementation of the mass penalty. Removing the penalty altogether (Table 13) leaves test MSE statistically unchanged (95% CI $[-0.0004, +0.0014]$) but lowers 1-JSD by 0.0075 (95% CI $[-0.0118, -0.0036]$). Hence, we retain the mass penalty in the main paper, as it keeps raw predictions, on which MSE is computed, close to valid distributions, and it improves distributional fit after renormalization.

Table 13: Stage 2 configuration ablation on the main panel (14 ATP waves, three repeated runs each; lasso; test split; Gemini-2.0-flash). Reported are means over the 42 runs. Agents is the mean number of agents with non-zero weight.

Configuration	MSE ↓	1-JSD ↑	EMD ↓	MCPA ↑	Agents
With mass penalty					
No uniform component (default)	0.0142	0.8300	**0.0657**	0.6764	105.7
+ Reference respondent	0.0130	0.8365	0.0699	0.6790	51.9
+ Intercept	**0.0128**	**0.8403**	0.0679	**0.6928**	60.4
Without mass penalty					
No uniform component	0.0146	0.8225	0.0672	0.6749	43.7
+ Reference respondent	0.0133	0.8381	0.0686	0.6782	44.3
+ Intercept	0.0132	0.8223	0.0771	0.6814	40.5

Anchors: reference respondent and intercept The default Stage 2 uses proxy agents to reconstruct observed response patterns of the target population. The design, albeit intuitive, requires proxy agents to account for the entire target distribution—every unit of predicted mass must come from some agent, including components that no agent expresses well, such as the uniform response mass of indifferent or inattentive respondents. Including anchors may aid Stage 2 selection, for they allow

the proxy agents to explain deviations from a reference distribution rather than levels, the same logic as the drift refit in Appendix D.6. We probe two *anchors* that are inspired by the uniform distribution. A *reference respondent* is an extra column with $d_{0,ik} = 1/K_i$, weighted and selected like any agent. Intuitively, it represents an indifferent respondent who answers a survey question by randomly picking an option. An *intercept*, in contrast, adds an unpenalized constant b to every option row, $\hat{r}_{ik} = \sum_j w_j d_{j,ik} + b$, so that $m_i(\mathbf{w}, b) = \sum_j w_j + K_i b$. Table 13 shows that on the main panel both anchors lower test MSE and raise 1-JSD with the mass penalty. The reference respondent also halves the selected ensemble, from 105.7 to 51.9 agents. However, neither anchor improves EMD. As no anchored configuration dominates across metrics, we keep the configuration without an anchor as the default.

D.4 Stress Test: Comparison with an SFT-Aligned Model under Topic Shift

P2P is designed as a plug-and-play preference alignment method that can be deployed on new populations with limited-data and limited-compute regimes. In the main paper, the baseline comparisons thus focus on prompting-based ensemble methods under the same constraints. As discussed in Appendix E, supervised finetuning (SFT) for a distribution-calibrated model offers an alternative approach under more lenient data and compute environments. An exemplar is Cao et al. [8], where SFT is used to train a predictive model based on pooled World Values Survey (WVS) data from 46 countries, and conditioned on locale to directly predict answer distribution for a given survey question. They further test the SFT-aligned model on a distinct unseen set of WVS questions along with other cross-country and cross-dataset validation tasks.

In this regard, our work differs from Cao et al. [8] not just in terms of training regimes, but also in key methodological premises. Specifically, we regard each survey panel as a benchmark for *a specific population preference on a given topic at a given moment in time*, where such preference is *gaugeable* through survey responses (cf. Figure 1, right panel). We do *not* assume that preferences on a topic domain can be reliably inferred if the training questions never probe that domain. For example, inferring political attitudes solely from non-political or religious items is a strong assumption for both humans and synthetic agents. Likewise, we do *not* assume preferences can carry over from one survey program to another, particularly when the surveys are conducted at different points in time and on different subpopulations of the same locale. Studying how aligned preferences generalize across topics, survey programs and over time is itself a key direction for future work, and will require both theoretical and empirical advances.

Nevertheless, Cao et al. [8]'s setup constitutes an interesting stress test for P2P outside our studied scenarios. We present in this section a head-to-head comparison with their method on their WVS arena. We follow their protocol: Q1 and Q2 for train/validation and Q3 as test, which induces a topic shift from general attitudinal & religious/ethical items (Q1+Q2) to political interest and culture items (Q3). In our language, this asks: if we align to a population on broad attitudinal dimensions (Q1+Q2), how well does the aligned ensemble generalize to a new unseen dimension (Q3: political-cultural questions)? As in Section D.2, we focus on the three representative locales: US (overlaps ATP), Great Britain (Western Europe) and Hong Kong SAR (East Asia). For each, we filter Cao et al. [8]'s data to that locale, train P2P (backend: gemini-2.0-flash; 300 endowments) on local Q1+Q2 (train+val), and evaluate on Q3. Table 14 shows the training schemes for P2P and Cao et al. [8].

Table 14: Different training regimes for P2P and [8].

Setup	Train+Val Entries (Q1 + Q2)	Test Entries (Q3)	Model backend
P2P (US)	185 (150 + 35)	59	gemini-2.0-flash
P2P (GB)	172 (140 + 32)	60	gemini-2.0-flash
P2P (HK)	185 (150 + 35)	59	gemini-2.0-flash
Cao et al. (C1: 46 countries)	8427 (6841 + 1586)	2719	Llama3-8B-Instruct

Table 15 presents the results on the performance comparison. In summary, P2P achieves 1-JSD on Q3 comparable to Cao et al. [8]'s SFT model and exceeds their majority-class prediction accuracy (MCPA), despite using far fewer training entries per locale and no fine-tuning (Table 14). Following

Table 15: P2P vs. Cao et al. [8] on WVS (Q3). Reported are mean ± sample SD over 3 repeated runs.

Setup	Method	Test MSE ↓	Test 1-JSD ↑	Test MCPA ↑	Test EMD ↓
P2P (US)	Lasso	0.0217 ± .0014	0.7473 ± .0031	0.6271 ± .0169	0.1174 ± .0051
P2P (US)	Enet	0.0218 ± .0019	0.7488 ± .0044	0.6384 ± .0098	0.1177 ± .0072
P2P (GB)	Lasso	0.0249 ± .0009	0.7416 ± .0055	0.4889 ± .0347	0.1296 ± .0045
P2P (GB)	Enet	0.0249 ± .0009	0.7463 ± .0067	0.5111 ± .0096	0.1294 ± .0048
P2P (HK)	Lasso	0.0284 ± .0023	0.7115 ± .0125	0.4972 ± .0353	0.1507 ± .0122
P2P (HK)	Enet	0.0290 ± .0023	0.7132 ± .0150	0.4859 ± .0427	0.1528 ± .0105
Cao et al.	SFT	–	0.777	0.43	–*

* EMD not reported due to incompatible definition in Cao et al. [8].

our earlier discussion, we interpret performance on Q3 as evidence that some preference structure transfers across topics, rather than as a requirement that a preference-alignment method must perfectly extrapolate to arbitrary unseen domains.

D.5 Subgroup Prediction: Comparison with SubPOP-FT

Section 4.4 demonstrates that a proxy basis generated for one population can be refit to another by refitting only its Stage 2 weights. The same principle applies to subgroups within a population. An expressive basis should represent each subgroup through different aggregation weights, without generating new agents or conditioning on demographics. We compare this approach with SubPOP-FT [64], a model fine-tuned specifically to predict subgroup response distributions. The methods differ in supervision: SubPOP-FT is fine-tuned on subgroup responses from other survey waves and predicts the target wave without seeing its responses, whereas P2P fits its aggregation weights on the training-question subgroup marginals of the target wave (see Table 20 for a broader comparison). Therefore, they serve complementary use cases, making SubPOP-FT an external reference rather than a competitor. We evaluate both methods on ATP and WVS.

ATP setup We use the OpinionQA subgroup benchmark of Suh et al. [64], rebuilding its subgroup targets from the official OpinionQA respondent release, with numerical agreement to within 5×10^{-9} across all 9,741 cells. Using the question splits of Table 1, we evaluate the 13 waves included in the benchmark, 22 subgroups and 208 held-out test questions. Following Suh et al. [64], refusals enter as zero entries on that question without removing the corresponding agents. We compute Wasserstein distance (WD), the benchmark's metric, alongside our three distributional metrics, with 95% intervals obtained by bootstrapping test questions (10,000 replicates). We report two agent panels with different levels of expressiveness: the gemini-2.0-flash bases of Table 1, and a second panel generated with GPT-4o answering the survey and learning attributes and GPT-4o-mini generating endowments.

Table 16 indicates that the default configuration yields the weakest performance on both panels, significantly underperforming SubPOP-FT on WD, MSE and 1-JSD. Adding an anchor closes most of the gap. With the reference respondent, P2P is statistically indistinguishable from SubPOP-FT on WD and MSE on both panels. On the GPT-4o/GPT-4o-mini panel, removing the mass penalty and adding an intercept matches SubPOP-FT on WD (95% CI $[-0.0092, +0.0095]$), MSE and 1-JSD, and exceeds it on MCPA (95% CI $[+0.002, +0.086]$).

WVS setup We use the WVS US subgroups common to Waves 6 and 7, with the three gemini-2.0-flash bases of Table 6, which were generated from W7 questions. We evaluate on 16 held-out W6 questions under three calibration regimes: (i) weights refit on W6 subgroup marginals; (ii) weights fit on W7 subgroup marginals and applied to W6; and (iii) weights fit on W7 population marginals and applied to every W6 subgroup. The third regime serves as a population-level reference, using no subgroup information. SubPOP-FT is applied zero-shot.

Table 17 shows that all P2P configurations outperform SubPOP-FT by a wide margin (test MSE 0.0088–0.0129 against 0.0249), and the other metrics agree: for the W6 refit without the mass penalty but with an intercept, 1-JSD is 0.841, EMD 0.098 and MCPA 0.644, against 0.737, 0.168 and 0.438. Notably, fitting on W7 subgroup marginals, without observing any W6 responses, retains nearly all of the performance of a direct W6 refit (0.0090 against 0.0088 with the reference respondent).

Table 16: Subgroup prediction on the OpinionQA benchmark of Suh et al. [64] (13 ATP waves, 22 subgroups, 208 test questions, 4,572 question–subgroup cells). P2P reports mean ± sample SD over three agent bases (lasso); SubPOP-FT uses the released adaptor. All four metrics, including MSE, are computed on P2P's clipped and renormalized predictions. WD is the normalized ordinal Wasserstein distance used by Suh et al. [64].

Configuration	WD ↓	MSE ↓	1-JSD ↑	MCPA ↑
P2P, gemini-2.0-flash panel (Table 1 bases)				
Mass penalty (default)	0.1231 ± .0014	0.0253 ± .0007	0.8257 ± .0028	0.6714 ± .0010
+ Reference respondent	0.1127 ± .0013	0.0192 ± .0004	0.8504 ± .0016	0.6817 ± .0070
+ Intercept	0.1155 ± .0013	0.0204 ± .0003	0.8477 ± .0013	0.6690 ± .0024
No mass penalty + intercept	0.1148 ± .0009	0.0187 ± .0003	0.8500 ± .0010	0.6897 ± .0054
P2P, GPT-4o/GPT-4o-mini panel				
Mass penalty (default)	0.1213 ± .0016	0.0269 ± .0014	0.8156 ± .0027	0.6820 ± .0137
+ Reference respondent	0.1100 ± .0033	0.0184 ± .0009	0.8529 ± .0034	0.7003 ± .0053
+ Intercept	0.1157 ± .0025	0.0217 ± .0009	0.8445 ± .0018	0.6852 ± .0141
No mass penalty + intercept	0.1069 ± .0019	0.0169 ± .0006	0.8584 ± .0027	0.7104 ± .0056
SubPOP-FT [64]	0.1066	0.0177	0.8663	0.6667

Subgroup-specific weights improve on population weights in every anchored configuration, though only modestly with the reference respondent (0.0088 against 0.0092). We read this as evidence of portability within WVS rather than of unconditional superiority: SubPOP-FT is trained on ATP, and extending it to WVS may require WVS-specific fine-tuning.

Table 17: Subgroup prediction on WVS (US; 16 held-out W6 questions, 10 subgroups common to W6 and W7). Reported is test MSE, mean ± sample SD over the three gemini-2.0-flash bases of Table 6 (lasso). SubPOP-FT is applied zero-shot.

Configuration	W6 subgroup refit	W7 subgroup fit → W6	W7 population fit → W6
Mass penalty (default)	0.0127 ± .0008	0.0129 ± .0001	0.0123 ± .0008
+ Reference respondent	0.0088 ± .0010	0.0090 ± .0002	0.0092 ± .0003
+ Intercept	0.0108 ± .0006	0.0112 ± .0001	0.0116 ± .0007
No mass penalty + intercept	0.0093 ± .0010	0.0096 ± .0002	0.0101 ± .0001
SubPOP-FT (zero-shot)		0.0249	

Connection to other sections of the appendices Three observations connect to our other ablation studies. First, subgroup prediction benefits from anchors far more than population prediction (cf. Appendix D.3.5). The reference respondent has a mean weight of 0.37–0.41 on OpinionQA/ATP and 0.35–0.46 on WVS, consistent with anchors letting the agents explain each subgroup's deviation from a reference distribution. Second, treating refusals as zero entries keeps agents with missing answers in the basis, qualifying for the first edge case of Appendix D.3.5: a question's mass then sums only over its answering agents—no longer equivalent to a single sum-to-one constraint on the weights—and the configuration without it performs best on the GPT-4o/GPT-4o-mini panel. Third, the model backend matters, echoing the backend ablation (Appendix D.3): with the same configuration, the GPT-4o/GPT-4o-mini panel reaches WD 0.1069 against 0.1148 for gemini-2.0-flash, though the difference is uneven across waves. Taken together, the comparison highlights two ways of exploiting survey data. When extensive subgroup-labelled data are available, supervised fine-tuning can amortize them into a specialized predictor. When target-population observations are limited but informative, P2P offers a portable alternative that reuses a response-expressive basis and updates only its aggregation weights. Cao et al. [8] remains the closer comparison for our setting, as it holds the survey program and target populations fixed and varies only the topic (Appendix D.4). We provide an extended comparison of the methods in Appendix E.3.

D.6 Use Case: Retrospective Imputation via Temporal Re-anchoring

When a new survey introduces questions absent from earlier waves, shared questions can calibrate a synthetic-agent basis to the historical population. The calibrated agent ensemble can then supply counterfactual response distributions for the previously unasked questions, extending series that would otherwise begin only at the later wave. This enables longitudinal and regression analyses that would otherwise suffer from missing data by design. We demonstrate this use case of `P2P` with Waves 6 and 7 of the WVS. The proxy agent pool is constructed using the more recent WVS W7 training questions (Appendix D.2), and re-anchored to the historical WVS W6 population using a limited number of shared questions. To make this otherwise unverifiable counterfactual testable, we adopt a masked missing-question design: we conceal the observed marginals of some questions that were actually asked, treat them as historically unasked, and reveal them only for evaluation.

Re-anchoring schemes We analyze two schemes requiring different levels of information:

- **Direct refit** uses only WVS W6 training marginals and predicts levels. Stage 2 is refitted to W6 as in the main pipeline.
- **Drift refit** uses both WVS W6 and W7 training marginals and predicts deltas. In practice, the W7 marginals enter every prediction with weight one, $\hat{r}_{ik}^{\mathrm{W6}} = r_{ik}^{\mathrm{W7}} + \sum_j w_j d_{j,ik}$, and the agent weights sum to zero, so that each prediction still sums to one.

We regard drift refit as more principled: predicting levels pressures the ensemble to reconstruct the full preference profile, including stable preferences it may fail to capture under limited data and expressiveness (where learnability, Definition 2.2, binds), whereas differencing frees the basis to focus on genuine context-dependent and drifting preferences. We fit the direct refit under three Stage 2 configurations (Appendix D.3.5) and the drift refit with ridge, lasso, and elastic net penalties. As a reference, we additionally introduce a *persistence* baseline, where the W7 marginals are carried back unchanged.

Evaluation We further report two errors capturing distinct reconstruction settings:

- **Masked-train MSE** evaluates questions whose W6 marginals are withheld from refitting but whose wording has contributed to constructing the W7 agents—the topic is covered by agent construction, but the historical marginals are unavailable.
- **Test MSE** evaluates the fixed W7-test questions, which contributed neither to agent construction nor regression fitting or tuning—agents constructed from an existing contemporary survey are used to reconstruct historical responses for newly introduced questions.

Protocol We use the original W7 split among 125 verified shared questions (90 train, 19 validation, and 16 test). Across 10 random splits and the three cached Table 2 bases, we conceal 18 train questions, re-anchor using nested subsets of the remaining 72 (5–72 questions), use validation only for hyperparameter selection, and reserve test questions for final evaluation. Results are shown in Table 18.

Results We make three observations based on Table 18:

- **Both schemes improve overall as more W6 calibration questions are supplied.** As the calibration budget increases from 5 to 72, MSE falls by 5.4–20.6% on test questions and by 9.9–18.9% on masked questions across configurations, though not always monotonically.
- **Direct refit test MSE falls with budget**, from 0.0122 to 0.00969 (for the configuration without mass penalty and uniform component), moving towards the original W7 error on the same shared test questions (0.00868). Masked-train MSE similarly drops from 0.0175 to 0.0142. Notably, both W6 refit schemes use far fewer training questions as anchors (capped at 72) than the W7 baseline (170).
- **Drift refit consistently outperforms persistence at every budget.** At 72 anchors, the ridge drift refit reduces MSE by 23.3% on masked-train questions, and 12.1% on test questions.

These results support using a synthetic agent ensemble for historical counterfactuals, especially via drift prediction, and offer an initial assessment of question-budget scaling.

Table 18: W6 reconstruction via the W7 proxy basis (WVS, US; three bases × ten random splits). Reported is mean MSE (in units of 10^{-3}) over the 30 basis–split cells, with SD in parentheses (for persistence, across the ten splits). Direct refits predict the W6 marginals with elastic net. Drift refits predict the W6−W7 change using W7 as the reference, with agent weights summing to zero. All fits use option-level weighting, with penalties chosen on 19 fixed validation questions. Persistence uses the W7 marginals unchanged. Bold indicates the lowest MSE per column and panel.

Configuration	Calibration anchors 5	10	20	40	60	72
Test questions (16)						
Persistence (W7 marginals)	3.64					
Direct refit, with mass penalty						
No uniform component (default)	11.64 (0.71)	11.56 (0.72)	11.28 (0.92)	10.93 (0.80)	11.19 (0.65)	11.01 (0.42)
+ Reference respondent	10.69 (1.24)	11.06 (0.78)	10.63 (1.05)	10.07 (1.17)	9.92 (1.04)	9.67 (1.15)
Direct refit, without mass penalty						
No uniform component	12.20 (1.66)	11.58 (1.66)	11.05 (1.10)	10.31 (0.80)	10.04 (0.75)	9.69 (0.60)
Drift refit, W7 reference						
Ridge	**3.42** (0.21)	**3.42** (0.23)	**3.39** (0.15)	**3.24** (0.13)	**3.23** (0.12)	3.20 (0.09)
Lasso	3.47 (0.18)	3.57 (0.20)	3.47 (0.27)	3.37 (0.21)	3.26 (0.15)	3.22 (0.14)
Elastic net	3.44 (0.18)	3.55 (0.20)	3.43 (0.24)	3.34 (0.20)	3.26 (0.14)	**3.18** (0.15)
Masked train questions (18)						
Persistence (W7 marginals)	3.01 (1.38)					
Direct refit, with mass penalty						
No uniform component (default)	16.98 (2.83)	16.84 (2.91)	15.86 (2.74)	15.48 (2.41)	15.52 (2.40)	15.31 (2.72)
+ Reference respondent	16.31 (2.66)	16.15 (2.59)	15.08 (2.59)	14.73 (2.52)	14.30 (2.53)	14.23 (2.80)
Direct refit, without mass penalty						
No uniform component	17.46 (4.03)	16.96 (3.80)	15.52 (2.52)	14.90 (2.37)	14.35 (2.33)	14.16 (2.48)
Drift refit, W7 reference						
Ridge	2.74 (1.14)	**2.56** (1.18)	**2.51** (1.19)	**2.46** (1.23)	**2.36** (1.20)	**2.31** (1.16)
Lasso	2.75 (1.14)	2.77 (1.22)	2.62 (1.24)	2.59 (1.29)	2.47 (1.24)	2.44 (1.24)
Elastic net	**2.69** (1.11)	2.74 (1.22)	2.61 (1.28)	2.56 (1.27)	2.45 (1.25)	2.39 (1.19)

E Extended Literature Review

E.1 Pluralistic Alignment

AI alignment refers to the process of ensuring that an AI system operates in accordance with human intentions and values, whether at an individual or aggregate level [30, 40]. As AI systems are adopted by increasingly diverse users, they must be designed to recognize and address a wider range of needs. This necessitates pluralistic systems capable of capturing and representing the diversity of human values and perspectives [62]. Pluralistic alignment therefore seeks to align models not with a single gold standard, but with a diverse range of user preferences across various attributes.

Most works on pluralistic alignment touch on three complementary ways, namely overton, steerable, and distributional pluralism, in which a single AI model or system can support diversity of views. In overton pluralism, the model aims to output a whole spectrum of reasonable responses [61]. Methods to achieve this include looking at alignment shifts to transform distributions toward overton coverage [37] or few-shot prompting from community-specialized LLMs [20]. In steerable pluralism, models are typically steered to adopt or favor particular perspectives or value attributes, often through few-shot examples [20, 3]. By contrast, our approach steers purely through attributes, with the goal of constructing diverse agents that embody different perspectives. In distributional pluralism, the distribution of the model over possible answers is intended to match that of some target population [62], often achieved through model fine-tuning on human responses [8, 64].

Another way to categorize pluralistic alignment work is by the amount of data they rely on. While most approaches in pluralistic alignment involve post-training, they range from resource-intensive approaches that fine-tune on large annotated corpora to lightweight methods that operate with minimal or no additional data. With sufficient data, specialized community models can be fine-tuned and subsequently combined using model merging techniques [68], few-shot prompting from community model outputs [20], or federated averaging [63] to scale with user diversity. If limited data are available, an alternative line of post-training approaches leverage in-context learning, where similarity-based retrieval [3] or group-informed retrieval [12] is performed to select few-shot examples that guide models toward pluralistic alignment. In the absence of detailed data from individual pluralistic users, a third post-training approach is inference-time conditioning. Inference-time conditioning uses structured prompts that encode sociodemographic traits or behavioral dispositions to condition an LLM to emulate the responses of a specific agent. These agents are also known as *endowments*, personas [10] or Silicon Samples [24, 4] and are lightweight to deploy and data-efficient. As inference-time conditioning is cheap and fast, they enable mixture or vector-based formulations of pluralistic alignment, which combine agents via multi-objective rewards to avoid averaging out minority perspectives [11, 19]. Therefore, `P2P` follows this line of work for pluralistic alignment. In the first step, `P2P` performs inference-time conditioning to generate a sufficient set of agents that adequately span the preference space. Thereafter, in the second step, it combines agents to achieve a compact yet pluralistic set of preferences that represent the target survey population.

Other related works on pluralistic alignment relate to the curation of datasets, which support research on pluralistic alignment [61, 71].

E.1.1 Attributes

Pluralistic alignment works that use inference-time conditioning often steer agents toward specific profiles using *attributes* that include moral and value dimensions [3], or broad morals, values, characteristics, and perspectives [61]. Attributes may also refer to demographic features, that are more tangible. Examples of these attributes include age, sex, education, income, and religion [10]. `P2P` uses the term *attributes* more broadly to encompass both moral and value dimensions and demographic features. Bundling these together allows `P2P` to define and condition LLM agents in a way that reflects both who they are and how they evaluate options.

E.2 Large Language Models to Emulate Human Preferences

Leveraging the ability of LLMs to emulate human or human-group responses with considerable fidelity, LLM alignment or personalization [66] has been largely focused on emulating human responses using various techniques. These include learning from feedback or reward models [6, 14], under distribution shifts [40, 36], finetuning [65], retrieval augmented generation [50], representation

learning [51], and using human-generated [24, 10, 4] prompts. However, when diverse responses need to be generated from many users or user groups instead of individuals, the data and time resources required for alignment grow linearly with the number of LLMs that need to be aligned. To reduce alignment costs, some works instead rely on AI-generated [56, 52, 21, 59] prompts, or alignment to multiple user groups [45, 70], even entire populations [8]. However, even the best models struggle with this task [8]. Furthermore, responses from LLMs have been found to have less variation than responses from real surveys [7], and LLMs have a tendency to respond in the middle category [67]. Consequently, it is imperative to address these limitations and devise data- and time-efficient approaches that enable LLMs to more accurately and reliably serve as proxies for survey populations. Our work contributes to this goal by introducing an active learning–inspired alignment framework that dynamically constructs a diverse agent basis under resource constraints and reconstructs population-level preferences via regression-based aggregation, without relying on ground-truth demographic profiles.

E.3 Comparison with Related Methods

Table 19 sets out how P2P differs from the methods most relevant to it. The key distinction is where the heterogeneity comes from and how it is adapted to a target population. Existing approaches either rely on externally supplied representations, such as demographic personas, retrieved perspectives or specialized community models, or learn implicit representations inside model or adapter parameters (GPO [70], Cao et al. [8], Suh et al. [64]). P2P instead actively constructs the heterogeneity required by the task at hand. Stage 1 searches for a response-expressive basis that is survey-specific (optimizing endowments based on survey question entropy) but population-independent (without observing target-population response patterns). Target-population calibration responses are only used in Stage 2 to estimate the target-specific basis composition. This separation is what enables a single basis to be reused across locales (Section 4.4).

The same comparison can be read as a question of where and how scarce human survey data should enter the alignment pipeline. Cao et al. [8] pool cross-country data to fine-tune a locale-conditioned predictor, and Suh et al. [64] use U.S. data to train a demographic specialist (SubPOP-FT). Krsteski et al. [35] divide survey data between response synthesis and post-hoc correction. P2P explores a further point in this design space: it leaves the backbone model unchanged and adapts only the inference procedure. Table 20 details the two fine-tuning approaches vis-à-vis P2P. Of these, Cao et al. [8] is the closer comparison, since it holds the survey program and target populations fixed and varies only the topic. We therefore use it for the stress test in Appendix D.4. Additional comparisons with Suh et al. [64] can be found in Appendix D.5.

E.4 Data Selection and Active Learning

To increase preference coverage while operating under budget constraints, our active endowment generation pipeline draws inspiration from machine learning paradigms, particularly active learning.

Traditional data selection methods aim to improve the data efficiency of supervised deep learning models by identifying the most informative or important training examples for generalization [16, 48]. Selection via Proxy (SVP) [16] uses a lightweight proxy model to estimate sample utility at a fraction of the cost, while GraNd and EL2N scores [48] identify important training examples early in training.

Similarly, deep active learning applies query strategies to iteratively select the most informative samples from a large pool of data to be added to the training dataset for retraining or fine-tuning [5, 27, 41]. Active learning strategies balance exploitation, focusing on uncertain or high-impact samples, with exploration, seeking diverse and novel samples to broaden coverage.

In our system, we apply this logic to guide both endowment generation and question patching. Specifically, low-entropy responses are treated as indicators of low coverage or model uncertainty. We refer to question patching as a boosting strategy—not in the ensemble learning sense, but to distinguish it from the main adaptive sampling process—since it targets underperforming questions by allocating additional sampling to enhance expressivity. Meanwhile, the exploitation–exploration mechanism used to sample across existing modes reflects the same active learning principle, but with a different emphasis: prioritizing high-utility regions while maintaining representational diversity. Together, these components form an active learning–inspired loop that incrementally expands coverage of the preference space while minimizing redundant agent generation.

Table 19: How P2P differs from related methods. Existing methods take heterogeneity from external representations or learn it in model parameters; P2P actively constructs it from the task at hand.

Method	Representation of heterogeneity	Supervision required	What is adapted
GPO [70]	In-context group preference examples	Cross-group preference data (meta-training) and target-group context examples	Preference module (meta-trained offline)
Modular Pluralism [20]	Specialized community models	Community-specific corpora	Community LMs (offline)
SPICA [12]	Retrieved group perspectives/scenarios	Group-labelled scenario resources	None (in-context)
PERSONA [10]	Synthetic demographic personas with LLM-enriched features	Census-derived demographic information	None (in-context)
Cao et al. [8]	Locale-conditioned predictor	$\sim$8.4K pooled response entries from 46 countries	Fine-tuned model weights
SubPOP-FT [64]	Subgroup-conditioned predictor	70K subgroup–response pairs from 53 ATP waves	LoRA weights
P2P	Response-expressive proxy basis	Target-population marginals on calibration questions	Proxy basis (from question text) and aggregation weights

Table 20: Comparison of Suh et al. [64] (SubPOP-FT), Cao et al. [8], and P2P.

	SubPOP-FT	Cao et al.	P2P
Target populations	22 U.S. demographic subgroups	46 country populations	Defined by the given survey
Applying to a new population	New training corpus and fine-tuning	Locale supplied as model input	Re-run the pipeline (new survey) or refit Stage 2 weights (same survey); locale conditioning not required (Section 4.4)
Adaptation	LoRA fine-tuning; GPU	Fine-tuning; GPU	Inference-time regression; CPU
What varies, training to test	Fielding period, respondents, topic	Topic only	Questions only
Temporal drift	Not modelled	Not modelled	Weights refit per survey occasion

E.5 Connections to Econometrics

P2P is related to three lines of econometric work. Synthetic control [1, 2] estimates the counterfactual outcome of a treated unit as a weighted combination of untreated donor units, with weights fit to reproduce the treated unit's pre-treatment outcomes and judged afterwards. In a similar spirit, P2P fits weights over proxy agents using training questions and evaluates them on held-out questions, although the agents are actively generated rather than taken from a pre-existing set. Indicator saturation [54, 31, 9] mitigates outliers and structural breaks in a time series by including a full set of dummy variables spanning the modeled time period and then selecting among them. P2P likewise generates more candidate agents than it retains, with Stage 1 aiming to span the latent preference space and Stage 2 selecting a compact ensemble from the pool. Viewed as a basis expansion whose size can grow, P2P's ensemble approach is akin to a sieve approximation [13] to the population response distribution. We leave a formal treatment of these connections to future work.

F Scope of Generalization

What generalization means in P2P P2P treats every survey as a snapshot of a particular set of preferences, representing a specific population, on a given topic, at a specific moment in time. It reconstructs these preferences only as identified through the survey—it does not recover broad-sense context-free preferences. As the right panel of Figure 1 illustrates, any observation set is a many-to-one, imperfect gauge of a lower-dimensional preference space, where more observations sharpen identification without removing the reduction. Our vantage is representation learning rather than psychological recovery: observed responses serve as anchors for learning an agent-ensemble representation of the human preferences that gave rise to them, and validity is only asserted where it can be tested, on held-out responses. Accordingly, generalization in P2P is to unseen questions within the same survey, with the population and topic scope held fixed. This scoping is consistent with established findings in social science: survey responses are shaped by the instrument itself (e.g., wording, format, and context) [57] and instruments sharing a topic label need not probe the same construct [28]. The econometrics of extrapolation makes the same point from the other side: transporting estimates to a new population requires explicit identifying assumptions [25], while extrapolation over time is fragile to shifts in the local data-generating process (LDGP) [15]. The ensemble P2P constructs recovers the LDGP of a given survey-population-moment.

Beyond the scope Applying the method beyond its generalization scope—by extending an agent ensemble, fit to a particular population and context, across topics, surveys, or time—calls for extra caution. We maintain healthy scepticism about methods claiming to recover general preferences, as that claim itself is a measurement claim, valid only if data supports the scope. Zero-shot transfer to a new topic, for instance, is a domain-shift problem, inherently harder than generalizing to questions within the same topic, which is why the comparison with Cao et al. [8] is framed as a stress test (Section D.4). For P2P, we make the distinction between the generated candidate pool (output of Stage 1, represented using endowments) and the final agent ensemble (output of Stage 2, represented using endowments and weights), and observe that the former can be used across populations and moments. Accordingly, the locale transfer (Section 4.4), the subgroup prediction (Section D.5) and the temporal re-anchoring experiment (Section D.6) are transfer learning in the standard sense: a candidate pool generated for a specific survey is reused by refitting its weights for a new target—the familiar pattern of a frozen representation with a target-supervised head.

G Limitations and Future Work

Question weighting and regression loss By design, P2P converts all multiple choice questions into binary questions before fitting a regression model. For example, a question with five options will turn into five questions with binary options. While this design ensures that regression can run smoothly with heterogeneous question types, it steers the model to prioritize fitting questions with more answer options, e.g., a 10-point Likert scale. Put differently, information contained in questions with fewer answer options, e.g., Yes/No, gets diluted after the binarization. Whether or not this design is desirable depends on the user's assumption of the relationship between preference signal and choice granularity. The current design implicitly posits a positive relationship between the two—i.e., the finer the scale, the better a signal on the latent preference. For other assumptions, question reweighting or format transformation need to be done before the regression step. P2P uses option-level weighting as the default, but also offers block weighting as an option.

Survey data in, text out The current implementation of P2P is limited to structured response settings: Given question and options, the aligned agent ensemble expresses its preference by offering a probability distribution over the given options. A key future research direction is to extend the use case to text completion, e.g., drafting resolutions, offering suggestions and answering freeform survey questions. How to aggregate individual agent textual responses into a representative aggregate response remains an under-explored area of research. Auction design [cf. 18] can be a promising aggregation algorithm in this case.

P2P **as a benchmark for model steerability** In this study our primary goal is understanding *preference reconstruction theory* and its implementation through the two-stage alignment framework. While our focus is on the design aspects of the system, we acknowledge that the backend model's steerability is also a determinant of P2P's empirical performance. The ablation study using different backends offers first insights on this aspect. It also indicates P2P's potential of being transformed into a benchmark pipeline for the study of model steerability in pluralistic alignment.

Theoretical guarantees and finite-sample behavior Our theoretical formulation characterizes preference reconstruction at the level of learnability and expressive feasibility, but it does not yet offer finite-sample error bounds or convergence guarantees for the full P2P pipeline. Deriving these theoretical insights is non-trivial as proxy agents are actively constructed using a battery of adaptive sampling strategies rather than sampled from a fixed, well-specified distribution, placing the procedure outside standard statistical learning settings. Our simulation studies offer controlled evidence that reconstruction improves with proxy-pool expressiveness, and our empirical ablations show stable performance under moderate pool sizes, but these results should be interpreted as empirical validation rather than formal guarantees. Developing error bounds that relate reconstruction quality to proxy expressiveness, calibration-target question shift, finite sample size, and regularization constitutes an important direction for future work.

Temporal Alignment and Non-stationary Preference Any method calibrated at a single moment risks becoming stale as attitudes and beliefs shift. This also affects P2P when considering using a fitted ensemble in future surveys. Our view is that synthetic agents should be periodically re-anchored using more recent, easier-to-collect data (e.g., shorter or cheaper surveys) and then tested on more demanding items. Architecturally, P2P is set up for this. The endowment generation and regression stages can be rerun on new waves, and one can track how learned weights and entropy patterns evolve, or when reconstruction quality begins to deteriorate. Appendix D.6 demonstrates such re-anchoring across two WVS waves. While P2P provides the machinery, systematic studies of preference shift and alignment across time and topics remain open.

Anchor-Centered Stage 2 The anchors of Appendix D.3.5 and the drift refit of Appendix D.6 inspire a more general redesign of Stage 2: fitting deviations from a reference distribution rather than response distributions directly. Given an anchor distribution a_{ik}, we can subtract it from both the targets and every agent's answers before fitting:

$$r_{ik} - a_{ik} = \sum_j w_j \left(d_{j,ik} - a_{ik}\right) + \varepsilon_{ik}, \text{ equivalently } \hat{r}_{ik} = \left(1 - \sum_j w_j\right) a_{ik} + \sum_j w_j \, d_{j,ik}. \quad (15)$$

As centered answers sum to zero within each question, any weights yield predictions that sum to one, removing the need for a mass penalty. The ℓ_1 penalty also becomes genuine selection. In the current design, weights summing to one fix $\|\mathbf{w}\|_1$ up to the negative weight, so the penalty mainly prices negative weight. With centering, shrinking the weights shrinks the prediction toward the anchor, a valid distribution. When no agents are selected, Stage 2 returns the anchor, while each selected agent explains a deviation from it. The anchor can be chosen to match the setting: the uniform distribution, recovering the reference respondent with an unpenalized weight $1 - \sum_j w_j$; the previous wave's marginals, recovering the drift refit when $\sum_j w_j = 0$; the overall population's marginals when predicting subgroups; or the mean of the agents' answers, so that Stage 2 shrinks toward the simple average. We leave the implementation and evaluation of this design to future work.

Short-cut Learning and Generalization Overfitting to survey artifacts is a key concern for *any* survey-based evaluation. This concern is closely related to what the robustness literature calls *shortcut learning* or *Clever Hans* behavior, where models exploit superficial regularities of a dataset instead of learning the underlying concept [22, 38]. Our empirical evaluation across 14 ATP waves is designed to mitigate the risk that our findings hinge on idiosyncrasies of a single wave, and the additional WVS experiments further mitigate the risk of ATP-specific shortcutting. While these steps serve as initial evidence against program-specific shortcutting, they do not constitute a complete solution to the generalization problem. A full multi-survey study (e.g., across GlobalOpinionQA and other polling corpora) is beyond the scope of this paper, but `P2P` 's modular pipeline readily supports this kind of analysis. Fundamentally, we see this as partly a survey design challenge: if we want synthetic agents to learn stable preferences rather than survey-specific shortcuts, future work may need survey instruments that are explicitly constructed with learnability and cross-topic generalization in mind. Moreover, we suggest a semantic question module for `P2P` to quantify the expressiveness of survey questions, which could serve as a parallel line of work to mitigate shortcutting on the model-design side. Taken together, these directions highlight an important and promising interdisciplinary research frontier in our pursuit of alignment science.

H Broader Impacts and Ethical Considerations

The work endeavors to lower the fixed cost of entry for preference alignment and computational social science research. By requiring only API inference, publicly available social survey datasets, and standard CPU-based regression, `P2P` enables researchers without access to GPU clusters, proprietary individual-level datasets, or sensitive demographic profiles to study population-level preference reconstruction. As a zero-demographic method, it also serves as a diagnostic tool for testing when finer demographic or locale conditioning is genuinely necessary for faithful emulation.

At the same time, the method relies on LLM backends that may propagate social biases, stereotypes, or culturally skewed response patterns, and some generated endowment profiles may contain opinionated or sensitive content. Our goal is methodological—studying preference reconstruction at the aggregate level rather than individual personalization—which reduces profiling risks. We see potential for positive impact on pluralistic alignment and principled use of AI in social science research, but stress that the agent ensemble is an auxiliary tool for understanding human choice patterns and should not replace humans in critical policy-making scenarios.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction accurately state the paper's main contributions and empirical scope: the preference reconstruction formulation, the two-stage P2P system, and evaluation on ATP and WVS, which are supported by Sections 2, 3 and 4 of the paper.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Appendix G discusses limitations including option-level weighting, finite-sample theoretical guarantees, entropy as a proxy for preference diversity, dependence on backend steerability, temporal non-stationary preferences, shortcut learning and extension beyond structured responses.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: Section 2 states the theoretical definitions, observations, and assumptions underlying preference reconstruction. Appendix A provides the formal proofs for the supporting propositions and observations used in the paper.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Sections 3 and 4, the appendices, and the supplemental material describe the data pipeline, data sources, train/validation/test splits, model backends, prompts, hyperparameters, baselines, and evaluation metrics needed to reproduce the main experimental results. The software used for the camera-ready analyses is available at `https://github.com/bingchen-wang/P2P`. Exact replication may be affected by provider-side nondeterminism in API-based LLM inference, which is outside the experimenter's control.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.

(c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The supplemental material provides the code and YAML config files for replicating the main experiments in the paper, and the `P2P` software, used for the camera-ready analyses, is available at `https://github.com/bingchen-wang/P2P`. Data sources and provenance are clearly documented in the paper, and the raw datasets are publicly available from their original providers.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Training and test details are clearly stated in Sections 3 and 4.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: The main results report means and sample standard deviations over repeated runs, computed under the same experimental condition, and the appendices report 95% bootstrap intervals for paired comparisons.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The paper reports monetary API costs for the main experiments and describes the compute requirements. We note that the experiments can be run with standard CPU resources and the runtime is dominated by API inference. Additional backend and implementation details are provided in the appendices and supplemental material.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We abide strictly by the NeurIPS Code of Ethics. Additionally, we provide a clear discussion on ethical considerations related to LLM bias, profiling risk, and appropriate use of synthetic agents in Appendix H.

   Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Appendix H provides a detailed discussion on the paper's broader impacts.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release a pre-trained model, image generators, scraped datasets, or other high-risk asset. The datasets used in the paper are publicly available and anonymized survey datasets.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The paper cites and credits the original sources of the datasets, benchmarks and models used in the experiments, including ATP/OpinionQA, WVS, PERSONA, and the SFT-based comparison studies. Data sources and provenance are clearly documented in the paper and supplemental material.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The paper releases the `P2P` software at `https://github.com/bingchen-wang/P2P` and documents the experimental and reproduction instructions. No new human-subject dataset or pretrained model is released.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not contain crowdsourcing experiments or research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not collect new human-subject data.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: LLMs are a core component of `P2P`, as the paper studies the use of LLM-based synthetic agents in emulating human population responses. In addition to the use of LLMs as synthetic respondents, they are used for attribute learning and endowment generation. Section 3 documents their roles in the pipeline, and the associated prompts are provided in the appendices.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.