Language Model Fine-Tuning on Scaled Survey Data for Predicting Distributions of Public Opinions

Anonymous ACL submission

Abstract

Large language models (LLMs) present novel opportunities in public opinion research by predicting survey responses in advance during the early stages of survey design. Prior methods steer LLMs via descriptions of subpopulations as LLMs' input prompt, yet such prompt engineering approaches have struggled to faithfully predict the distribution of survey responses from human subjects. In this work, we propose directly fine-tuning LLMs to predict response distributions by leveraging unique structural characteristics of survey data. To enable fine-tuning, we curate SubPOP, a significantly scaled dataset of 3,362 questions and 70K subpopulation-response pairs from well-established public opinion surveys. We show that fine-tuning on SubPOP greatly improves the match between LLM predictions and human responses across various subpopulations, reducing the discrepancy in distribution over option choices by up to 46% compared to baselines, and achieves strong generalization to out-of-distribution data. Our findings highlight the potential of survey-based fine-tuning to improve predictions about opinions of real-world populations and therefore enable more efficient survey designs.

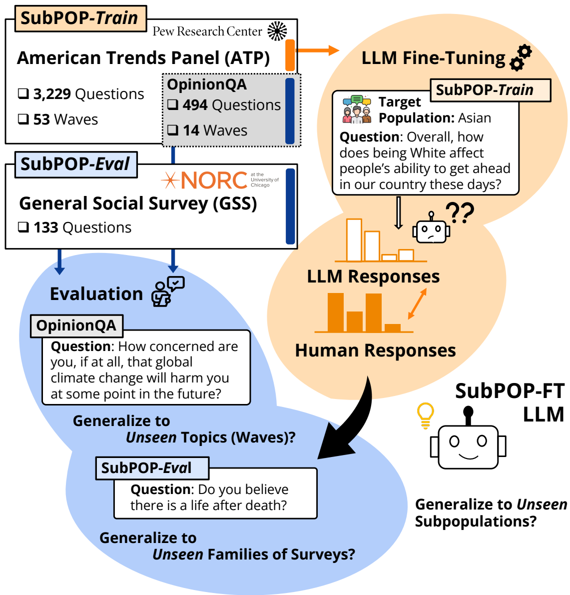

Figure 1: Illustration of our method and SubPOP. We collect survey data from two survey families—ATP from Pew Research (Center, 2018) (forming SubPOP-Train) and GSS from NORC (Davern et al., 2024) (forming SubPOP-Eval). LLMs are fine-tuned on SubPOP-Train and evaluated on both OpinionQA (Santurkar et al., 2023) and SubPOP-Eval to assess generalization of distributional opinion prediction across unseen subpopulations, topics, and survey families.

1 Introduction

Surveys provide an essential tool for probing public opinions on societal issues, especially as opinions vary over time and across subpopulations. However, surveys are also costly, time-consuming, and require careful calibration to mitigate non-response and sampling biases (Choi and Pak, 2004; Bethlehem, 2010). Recent work suggests that large language models (LLMs) can assist public opinion studies by predicting survey responses across different subpopulations, explored in both social science (Argyle et al., 2023; Bail, 2024; Ashokkumar et al., 2024; Manning et al., 2024), and NLP (Santurkar et al., 2023; Chu et al., 2023; Moon et al., 2024; Hämäläinen et al., 2023; Chiang and Lee, 2023). Such capabilities could substantially enhance the survey development process– not as a replacement for human participants but as a tool to complement various phases, *e.g.* pilot testing (Grossmann et al., 2023; Ziems et al., 2024; Rothschild et al., 2024; Dillion et al., 2023; Learner, 2024).

Prior work in steering language models, *i.e.* conditioning models to reflect the opinions of a specific subpopulation, has primarily investigated different prompt engineering techniques (Santurkar et al., 2023; Moon et al., 2024; Park et al., 2024a). However, prompting alone has shown limited success in generating completions that accurately reflect the distributions of survey responses collected from human subjects. Off-the-shelf LLMs (Achiam et al., 2023; Dubey et al., 2024; Jiang et al., 2023) have shown to mirror the opinions of certain US subpopulations such as the wealthy and educated (Santurkar et al., 2023; Gallegos et al., 2024; Deshpande et al., 2023; Kim and Lee, 2023), while generating stereotypical or biased predictions of underrepresented groups (Cheng et al., 2023b,a; Wang et al., 2024). Furthermore, these models often fail to capture the diversity of human opinions within a subpopulation (Kapania et al., 2024; Park et al., 2024b). While fine-tuning presents opportunities to address these limitations (Chu et al., 2023; He et al., 2024), existing methods fail to train models that accurately predict opinion distributions across (1) diverse subpopulations and (2) various survey question topics.

The present work. We propose fine-tuning LLMs on large collections of data from cross-sectional public opinion surveys, consisting of questions about diverse topics and full distributions of responses from each subpopulation defined by demographic and ideological traits. By casting pairs of (subpopulation, survey question) as input prompts, we train the LLM to align its response distribution against that of human subjects in a supervised manner. We posit that survey data is particularly well-suited for training LLMs since: (1) We can construct clear **subpopulation-response pairs** as data samples from which models learn associations between group identities and expressed opinions, which are typically rare in language models' pre-training corpora, (2) Large-scale opinion polls are carefully designed and calibrated (*e.g.* using post-stratification) to collect **representative** human responses, even for minority groups that have high empirical variance, (3) We can enable LLMs to capture subpopulation opinions as **distributions over multiple options** using a training objective that explicitly matches model predictions against response distributions of human subjects.

Training on public opinion survey data has remained under-explored due to the limited availability of structured survey datasets. To this end, we curate and release `SubPOP` (**Sub**population-level **P**ublic **O**pinion **P**rediction), a dataset of 70K subpopulation-response distribution pairs ($6.5\times$ larger compared to previous datasets). We show that fine-tuning LLMs on `SubPOP` significantly improves the distributional match between LLM generated and human responses. Additionally, the improvements strongly generalize to *unseen* subpopulations, survey waves, and survey families, *i.e.* surveys administered by different organizations. In particular, we observe that our approach addresses prior limitations in approximating opinion distributions of diverse subpopulations, including minority groups.

Our contributions are summarized as follows:

- We show that training LLMs on response distributions from survey data significantly improves their ability to predict the opinions of subpopulations, reducing the Wasserstein distance between model-predicted and ground-truth distributions by 32-46% compared to top-performing baselines. (Section 4.2)
- We show that the performance of the fine-tuned LLMs strongly generalizes to out-of-distribution data, including unseen demographic groups, new survey waves, and different survey families. (Section 4.2 and Section 4.3)
- We release `SubPOP`, a curated and pre-processed dataset of public opinion survey results that is $6.5\times$ larger than existing datasets, enabling fine-tuning at scale.

2 Related Work

Predicting Human Opinions via LLMs. Prior work has explored various prompt engineering approaches for steering LLM responses: earlier work use rule-based prompts that incorporate demographic profiles of individuals or populations, or few-shot examples of survey question-response (Hwang et al., 2023; Simmons, 2022; Santurkar et al., 2023; Dominguez-Olmedo et al., 2023). Recent work explore prompting LLMs with open-ended text, including interview transcripts (Park et al., 2024a), personal narratives (Moon et al., 2024), or LLM-refined prompts (Kim and Yang, 2024; Sun et al., 2024). Our proposed method of fine-tuning language models with survey response data is complementary to improvements in prompt engineering, because for prompt engineering improved prompts facilitate conditioning on the target group but in our approach LLMs are directly guided to use the target group label for opinion prediction. In this work, we also demonstrate that our fine-tuned models can exhibit significant improvements in matching the response distributions of humans without elaborate prompt engineering methods.

Other work (Chu et al., 2023; He et al., 2024; Feng et al., 2024) fine-tune language models on text corpora from specific communities (*e.g.*, Reddit) to infer the most popular response or response distribution for a given survey question. While this approach benefits from large-scale and continuously updated text corpora, it struggles with disproportionate representation and lacks comprehensive coverage of diverse subpopulations. An alternative approach (Zhao et al., 2023; Li et al., 2023, 2024) directly trains on survey data, with (Zhao et al., 2023) applying meta-learning to predict opinions of unseen groups and (Li et al., 2024) fine-tuning on cross-cultural survey responses to predict the most popular response. However, optimizing for the most popular response discards distributional information, and our experiments (Appendix C.1) show that this exacerbates distribution mismatch.

Datasets for LLM-based Opinion Prediction. Several research institutions conduct large-scale public opinion polls and release data from those surveys. Important examples include Pew Research Center's American Trends Panel (ATP), which consists of multiple waves of cross-sectional surveys on different topics, and the General Social Survey (GSS) from the NORC at the University of Chicago (Davern et al., 2024). Existing datasets have curated such data for evaluating LLM-based opinion predictions, including OpinionQA (Santurkar et al., 2023), a subset of ATP survey waves containing about 500 questions on contentious social topics. While OpinionQA is widely used in prior work (He et al., 2024; Zhao et al., 2023; Li et al., 2023, 2024), we find its total number of questions limited in scale for fine-tuning LLMs and instead use this dataset for evaluation. We further collect an extended set of survey data from ATP waves not included in OpinionQA, as well as from GSS to curate SubPOP.

Other datasets, such as GlobalOpinionQA (Durmus et al., 2023)—derived from the World Values Survey (World Values Survey, 2022) and the Pew Global Attitudes Survey (Pew Research Center, 2024)—and the PRISM dataset (Kirk et al., 2024) investigates how language models align with opinions from populations across the globe and different cultures. In our work, we focus on surveys conducted in the U.S. and target U.S. subpopulations as an initial demonstration of our approach's empirical validity. However, our proposed method for fine-tuning language models applies to any survey dataset with distributional information about subpopulation responses.

Pluralistic Alignment of LLMs. Recent literature on pluralistic and distributional alignment target a similar yet different problem in fine-tuning LLMs (Chakraborty et al., 2024; Melnyk et al., 2024; Poddar et al., 2024; Siththaranjan et al., 2023; Yao et al., 2024; Sorensen et al., 2024; Lake et al., 2024; Chen et al., 2024; Jiang et al., 2024). While this line of work shares a similar goal as ours in training models to reflect on opinions (and preferences) of diverse subpopulations, most work differ from ours in that they operate in the context of training against *pair-wise* preference orderings between alternative language model completions, extending the Bradley-Terry-Luce model (Rajkumar and Agarwal, 2014; Ouyang et al., 2022; Rafailov et al., 2024) or investigating alternative models to account for diverging preference orderings across populations. In contrast, our work trains the model to directly predict the opinion distributions of human subpopulations, where accurately matching distributions across a large variety of subpopulations is of paramount interest. Our work additionally focuses on the particular context of estimating human opinions about societal issues—the objective of public opinion research—which enables relatively straightforward supervised training on openly available, structured survey data as presented by SubPOP.

3 Methods

3.1 Matching between Model and Human Response Distributions

Our goal is to fine-tune an LLM to predict the distribution of responses for a multiple-choice question, conditioned on descriptions of a human subpopulation we want to simulate, typically a specific demographic or ideological subgroup. Consider the example in Figure 2: the question asks, "What do you think the chances are these days that a woman won't get a job or promotion while an equally or less qualified man gets one instead?" The available responses are: *A. Very likely, B. Somewhat likely, C. Not very likely, D. Very unlikely, and E. Refused*. In this case, the LLM will output a probability for each of the tokens corresponding to the choices A through E, thereby generating a complete response distribution that we aim to align with the true distribution observed in survey data.

Formally, let $q \in Q$ be a question, $g \in G$ be a subpopulation, and $\mathcal{A}_q$ be the set of possible choices for question q. An LLM with parameters θ produces

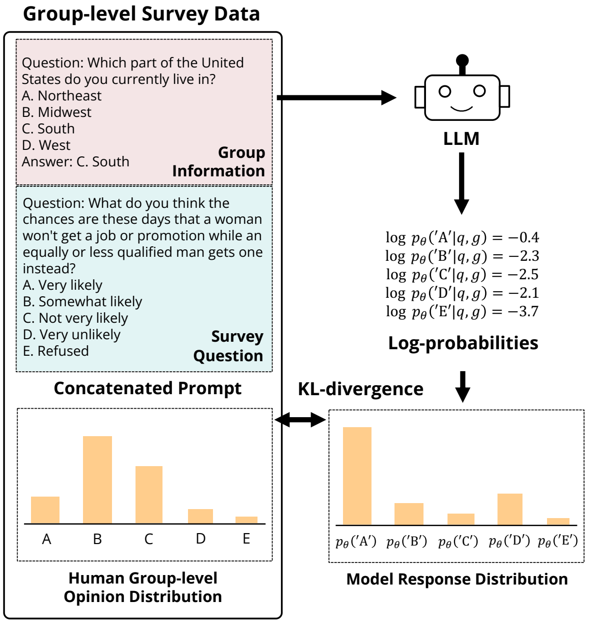

Figure 2: Proposed supervised fine-tuning setup with a survey response dataset such as SubPOP. Survey data is 3-tuple of a survey question, target subpopulation information, and the observed human opinion distribution (*i.e.* how subjects in the group responded to the given question). The training objective, $\mathcal{L}(\theta)$, is a forward KL divergence loss on language model predicted distribution of question option likelihoods; our loss guides the model predictions to match the response distribution of the specified human subpopulation.

a conditional probability distribution $p_\theta(\mathcal{A}_q \,|\, q,g)$. We fine-tune this model so that its predicted distribution for each (q,g) mirrors the human response distribution $p_H(\mathcal{A}_q \,|\, q,g)$ collected from real survey data.

To accomplish this, we apply LoRA fine-tuning (Hu et al., 2021) and use the forward Kullback–Leibler (KL) divergence as our loss. Concretely, if $p_H(\mathcal{A}_q \mid q,g)$ represents the group-level empirical distribution of human opinions and $p_\theta(\mathcal{A}_q \mid q,g)$ represents the model's predicted distribution, our training objective is:

$$\mathcal{L}(\theta) = \mathbb{E}_{q,g}\left[D_{\mathrm{KL}}\left(p_H(\mathcal{A}_q \,|\, q,g) \,\middle\|\, p_\theta(\mathcal{A}_q \,|\, q,g)\right)\right],$$

where D_{KL} denotes the KL divergence. In the example shown in Figure 2, the model is trained to reduce the KL divergence between the target (survey-based) distribution over $\{A, B, C, D, E\}$ and its predicted distribution for the subpopulation living in the Southern United States.

We choose forward KL (i.e., $\mathrm{KL}(p_H \,\|\, p_\theta)$) since it is sensitive to cases where p_H assigns high probability but p_θ does not, naturally encouraging the model to *cover* the real distribution. This property aligns with standard maximum-likelihood training, where the model is penalized for underestimating any response that is frequent in the data. In other words, if many participants in group g choose option "A" for question q, then the model probability on "A" should be correspondingly high.

Instead of explicitly modeling the group response distribution as $p_H(\mathcal{A}_q|q,\, g)$, one could do two alternatives. (1) One-hot encoding: this approach (Li et al., 2024) approximates the distribution by a one-hot vector, assigning a value of one to the most probable option and zero elsewhere. (2) Data augmentation by response frequency: this approach (Zhao et al., 2023) expands the dataset by replicating question-choice pairs in proportion to their observed frequency. We adopt the explicit distribution modeling in our main experiments because it directly encodes the distributional information without requiring discrete sampling or replicating data points. This avoids potential quantization errors introduced by binning continuous values and reduces the total amount of data needed. A detailed comparison of these approaches is provided in Section C.1.

3.2 SubPOP: a Comprehensive Survey Dataset to Fine-tune and Evaluate LLMs

OpinionQA (Santurkar et al., 2023) is a widely used dataset for fine-tuning and evaluating large language models (LLMs) on opinion prediction, containing roughly 500 questions drawn from 14 ATP (American Trends Panel) waves (Center, 2018). Although valuable, it faces two important limitations: (1) Limited thematic diversity—for instance, wave 26 focuses narrowly on firearms. (2) Reliance on a single survey family (ATP), which risks overfitting to a particular style of questions and hampers out-of-distribution evaluation on other sources (e.g., GSS).

To address these limitations, we introduce a new dataset, SubPOP, that broadens both the thematic and institutional scope of opinion prediction data. For training, SubPOP comprises 3,229 multiple-choice questions drawn from ATP waves 61–132. We exclude waves included in OpinionQA to assess whether an LLM fine-tuned with SubPOP can generalize to unseen subject areas. For evaluation, SubPOP includes 133 multiple-choice questions from the General Social Survey (GSS) (Davern et al., 2024), serving as an out-of-distribution benchmark. This expanded collection not only broadens the range of topics beyond OpinionQA's initial 500 questions, but also enables evaluation on surveys created and administered by different institutions (Pew Research Center ATP vs. NORC-Chicago GSS). Dataset curation and refinement pipeline is available in Appendix A.

3.3 Evaluation Metric

We use Wasserstein distance (WD) to quantify how closely the model's predicted opinion distribution matches human survey data (Santurkar et al., 2023; Moon et al., 2024; Meister et al., 2024; Zhao et al., 2023). Formally, for a group g representing some subpopulation and a question q WD is defined as $\mathcal{WD}_\theta(q, g) = \mathcal{WD}(p_H(\mathcal{A}_q|q, g), p_\theta(\mathcal{A}_q|q, g))$. Please refer to Appendix B for the exact formula of WD metric.

Some prior work utilizes one-hot accuracy (Feng et al., 2024; Li et al., 2023) as an evaluation metric. However, one-hot accuracy has a notable drawback for the response distribution prediction task. One-hot accuracy only verifies whether the top-predicted choice matches the top human response, thereby discarding distribution information. In contrast, WD accounts for partial overlaps among the categories and reflects the 'cost' of shifting probability mass, providing a more nuanced assessment of distribution discrepancy. Consider the example question provided in Figure 2, where the human response distribution indicates that option B ("Somewhat likely") is the most probable. Now consider two cases in which the model incorrectly predicts the top choice. In the first case, the model assigns a high probability to option A ("Very likely"), while in the second case, it assigns a high probability to option D. Although one-hot accuracy would treat both predictions equally as errors, WD differentiates between them by accounting for the ordinal relationship among the options, penalizing the second prediction more heavily for its larger deviation from the true distribution.

4 Experiments

4.1 Bounds of WD and Baselines

In this section, we describe the lower/upper bounds and two baseline methods against which we compare our method.

Lower and upper bounds. We use a uniform distribution over all available choices to establish an upper bound of the WD between a predicted and the target response distribution. To compute a lower bound, we sample a group of human respondents from the original human respondents to calculate the WD between the two, and perform bootstrapping to obtain a robust estimate. This lower bound captures the intrinsic variance arising from the respondent sampling process in opinion surveys.

Baselines. We compare our approach with two baseline methods: prompting and Modular Pluralism (Feng et al., 2024). For prompting, we consider both zero-shot and few-shot methods. In zero-shot prompting, we steer the LLM using demographic prompt formats. Specifically, we employ three different formats following Santurkar et al. (2023): `QA`, `BIO`, and `PORTRAY`. For instance, to condition the LLM to a person living in the South of the US, the `QA` format uses a question-answer format as illustrated in Figure 2; the `BIO` format conditions the model with a first-person narrative such as "I currently reside in the South."; and the `PORTRAY` format uses a third-person narrative like "Answer the following question as if you currently reside in the South.".

Few-shot prompting augments the prompt with a few examples of question-response distribution pairs alongside the demographic label (Hwang et al., 2023). In particular, we select the top five few-shot examples from the `SubPOP` training set based on cosine similarity computed by the embedding model. In our experiments, we represent the response distribution in JSON format and require the model to output its prediction in the same JSON format, following the approach in Meister et al. (2024).

Modular pluralism (Feng et al., 2024) fine-tunes multiple LLMs on distinct datasets to capture the viewpoints of different communities (Feng et al., 2023). For a given question, each fine-tuned LLM generates an opinion that reflects the perspective of the community it represents, and a separate black-box LLM aggregates these outputs to produce the final distributional response. Detailed implementation of the lower/upper bounds and the baselines is provided in Appendix D.

4.2 Generalization to Unseen Topics and Survey Families

In this section, we assess the ability of our fine-tuned LLMs to generalize to unseen data—both in terms of new topics and entirely different survey families. To evaluate these aspects, we use OpinionQA to measure generalization to unseen topics, and

Table 1: Evaluation on OpinionQA and the SubPOP evaluation set (SubPOP-Eval) for 22 subpopulations following (Santurkar et al., 2023). We compute the WD by averaging over all questions and subpopulations. Lower and upper bounds of performance give guidance on how each method performs. For Modular Pluralism, we provide an error rate of one-hot prediction (†) (Section 3.3) which was used in the original paper.

Method	OpinionQA				SubPOP-Eval			
	Llama-2-7B	Llama-2-13B	Mistral-7B	Llama-3-70B	Llama-2-7B	Llama-2-13B	Mistral-7B	Llama-3-70B
Upper bound (Unif.)	0.178				0.208			
Lower bound (Human)	0.031				0.033			
Zero-shot prompt (QA)	0.173	0.170	0.153	0.138	0.206	0.196	0.187	0.160
Zero-shot prompt (BIO)	0.193	0.183	0.162	0.143	0.221	0.212	0.202	0.175
Zero-shot prompt (PORTRAY)	0.195	0.207	0.158	0.209	0.212	0.242	0.194	0.247
Few-shot prompt	0.186	0.175	0.174	0.166	0.217	0.194	0.175	0.182
Modular Pluralism	0.285 († 55.6%)				0.279 († 55.2%)			
Ours (SubPOP-FT)	0.106	0.102	0.096	0.094	0.121	0.113	0.115	0.096

SubPOP-Eval to test generalization to a different survey family.

We fine-tune four LLMs (Llama-2-7B, Llama-2-13B, Mistral-7B, and Llama-3-70B) on SubPOP-Train. We opt for pretrained LLMs rather than instruction-following models, as previous work has shown that pretrained models perform better on this task (Moon et al., 2024). A detailed comparison between these model types is provided in Appendix C.2.

Table 1 reports the average WD metrics computed over all demographic groups and survey questions, comparing our fine-tuned models against various baseline approaches.

Summary of Results. Our experiments show that fine-tuning on SubPOP-Train significantly outperforms all other methods, yielding a 32–46% reduction in WD on OpinionQA and a 39–42% reduction on SubPOP-Eval compared to the best baselines. Notably, SubPOP-Train is based on ATP data, while SubPOP-Eval is derived from GSS surveys—two distinct survey families that can differ in respondent pools, calibration techniques, and other methodological factors, leading to non-trivial distribution shifts despite both being representative of the US population. Furthermore, our fine-grained analyses at the wave level (see Appendix E) confirm that these trends persist even at more detailed levels of evaluation.

Comparison to Zero- and Few-Shot Prompting. We first compare the performance of prompting methods with our approach. Zero-shot prompting results in only modest WD improvements over the upper bound, with the largest gain observed for Llama-3-70B and negligible improvements for Llama-2-7B. Even when using few-shot prompting—where five example question-response distribution pairs are provided—the performance gains remain minimal. This may be partly due to an under-optimized prompt format (*e.g.* requiring JSON output) and the inherent sensitivity of language models to prompt formatting (Sclar et al., 2023; Anagnostidis and Bulian, 2024). These findings underscore the need for methods, such as fine-tuning, that enable relatively reliable predictions of opinion distributions.

Comparison to Modular Pluralism. Modular Pluralism improves one-hot accuracy, reducing prediction error from 72.7% (zero-shot prompting) to 55.6% on OpinionQA, but underperforms in matching the full distribution of option choices, measured as WD. This discrepancy in performance highlights the limitations of methods that train LLMs to identify only the most probable response rather than modeling the entire distribution of responses. Opinions are inherently distributed: even within a particular subpopulation such as a single demographic subgroup, distribution of opinions cannot be captured as a single most likely response. Moreover, instruction-tuned models that serve as a black-box LLM tend to assign high probabilities on only specific tokens (Lin et al., 2022; Kadavath et al., 2022; Achiam et al., 2023), further pushing the generated distribution away from the human distribution.

4.3 Generalization across Target Subpopulations

Here we report two key observations: (1) prediction performance improves consistently across most subpopulations represented in the fine-tuning data, and (2) the LLMs fine-tuned on SubPOP-Train generalize well to subpopulations that were not included during fine-tuning.

Consistent Performance Improvements over Subpopulations. Figure 3 shows the per-group WD on the OpinionQA evaluation for Llama-2-7B, comparing our fine-tuning approach with zero-shot

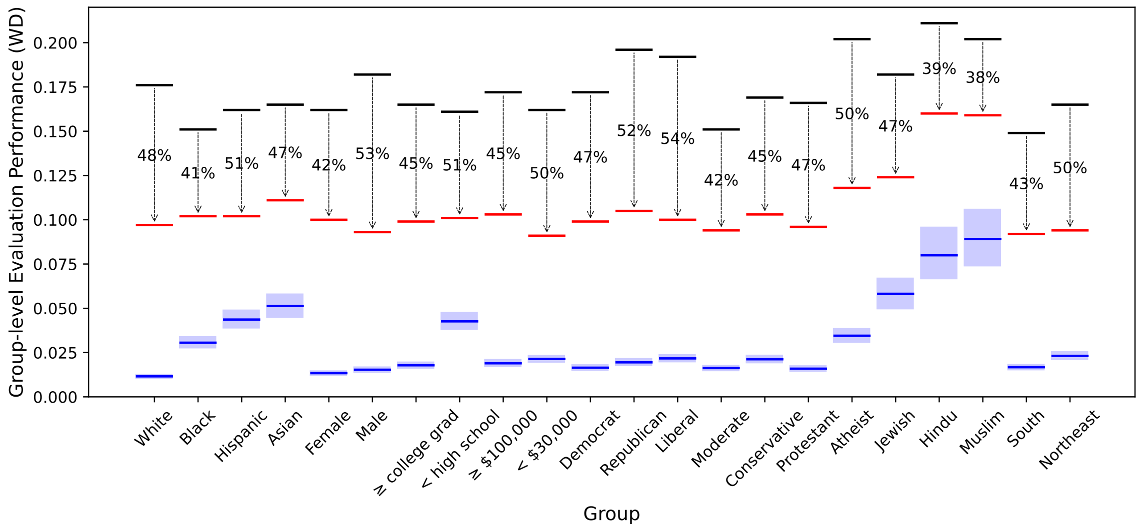

Figure 3: Per-group evaluation performance of our model Llama-2-7B-SubPOP-FT (red lines) on OpinionQA. For comparison, the results from zero-shot QA prompting (black lines) and the lower bound (blue lines) are presented. We observe that the relative improvement, measuring how much of the gap between zero-shot prompting and the lower bound has been closed, remains consistent across subpopulations. Shaded blue regions represent the 95% confidence interval of the lower-bound estimation for each group. Per-group results for other models (Table 7) and the results on SubPOP evaluation set (Table 8) are available in Appendix E.

prompting and the empirical WD lower bound. To evaluate the consistency of performance gains, we calculate the *relative improvement* for each subpopulation as how much of the gap between zero-shot prompting and the empirical lower bound is reduced after fine-tuning. This measure allows us to account for varying lower bounds across subpopulations: since some groups have fewer respondents, there is greater uncertainty in their reported distribution in the survey data and greater variance between the original sample and bootstrap samples.

With the exception of two of the smallest groups (Hindu and Muslim), all subgroups demonstrate a large and consistent relative improvement after fine-tuning, ranging from 40%–54%. Including all groups, the average relative improvement is 46.7%, with a standard deviation of 4.4%. This consistency confirms that our fine-tuning approach delivers balanced performance gains without disproportionately favoring any particular demographic subgroup. We hypothesize that the consistent gains over groups largely stem from our dataset design, which allocates an equal number of training samples to each group. By ensuring uniformly distributed data points across subpopulations, the model captures sufficient subgroup-specific signals, ultimately leading to consistent performance improvements.

Generalization on Unseen Subpopulations. We further investigate how models fine-tuned with our approach and SubPOP might show generalization to subpopulations that were not represented in the training data, a circumstance that can commonly occur when such fine-tuned LLMs are deployed for use in assisting survey design. For this evaluation, we benchmark our methods against a zero-shot prompting baseline. Specifically, we evaluate our model, which is fine-tuned on 22 subpopulations provided in SubPOP-Train, on a set of 38 subpopulations in OpinionQA that were not included in fine-tuning. This experiment not only checks generalization to unseen subpopulations, but also involves unseen survey questions, providing a robust assessment of the model capability for generalization to OoD data.

As shown in Table 2, our model achieves a strong reduction in WD even for unseen subpopulations, indicating that the model can be steered by demographic prompts beyond the seen subpopulations in training. Interestingly, although SubPOP-Train does not contain any data with opinion distributions of particular age groups (*e.g.* subjects of age 18-29 or those of age 65+), the average relative improvement is 44.7%, which is compatible with the average relative improvement for seen subpopulations.

For other traits such as education level and political ideology in Table 2, the relative improvements for unseen subpopulations is comparable with the relative improvements for seen subpopulations. We provide results for other unseen subpopulations in Table 6. For most of unseen subpopulations, our methods achieve comparable relative improvements. These findings show that our fine-tuning approach effectively steers the model with conditioning prompts and robustly generalizes to a wide range of

Table 2: Per-group evaluation performance of Llama-2-7B-SubPOP-FT (Ours) on OpinionQA. We report the lower bound, WD for zero-shot prompting, WD for Llama-2-7B-SubPOP-FT, and the relative improvement. Rows highlighted in blue represent subpopulations included during fine-tuning, while uncolored rows correspond to subpopulations that were unseen during fine-tuning.

Group	Lower Bound	Zero Shot	Ours	Relative Improvement (%)
Age: 18-29	0.023	0.185	0.096	54.8
Age: 30-49	0.014	0.151	0.093	42.4
Age: 50-64	0.014	0.154	0.101	37.7
Age: 65+	0.013	0.195	0.115	43.8
Less than high school	0.043	0.161	0.101	45.4
High school graduate	0.017	0.144	0.092	41.3
Some college, no degree	0.018	0.144	0.093	40.5
Associate's degree	0.026	0.159	0.098	45.5
College grad	0.018	0.165	0.099	51.2
Postgraduate	0.015	0.174	0.106	42.6
Very conservative	0.026	0.208	0.107	55.5
Conservative	0.021	0.191	0.110	44.7
Moderate	0.018	0.184	0.120	42.1
Liberal	0.018	0.224	0.102	54.2
Very liberal	0.025	0.202	0.111	51.4

subpopulations. The further analysis on this result is available in Appendix C.3.

4.4 Effect of Scaling the Dataset

In this section, we examine performance scales with training dataset size. We randomly sample subsets containing 25%, 50%, 75%, and 87.5% of the full SubPOP training set and evaluate three models—Llama-2-7B, Llama-2-13B, and Mistral-7B—on OpinionQA. As shown in Figure 4, we observe diminishing marginal returns, as is typical with fine-tuning; for example, after training on a random 25%, the models reach 72%-78% of the total improvement they achieve after fine-tuning on all of SubPOP-train. However, what is interesting is that performance does not entirely plateau. Instead, it continues to improve as we further increase the training data from 25% to 100%. We fit linear trend lines (dotted in Figure 4) to the results and observe that the slopes are similar for each model. This suggests that the rate of improvement—reflected by the slope in the power-law relationship—is intrinsic to the data and task rather than to the specific model architecture. In other words, LLMs exhibit comparable data efficiency, with performance gains that are fundamentally tied to dataset size rather than model-specific factors.

Using these trend lines, we can estimate the amount of fine-tuning data required to reach a target performance. For instance, we estimate that fine-tuning Mistral-7B on a dataset 25 times larger than the current SubPOP training set would yield a WD value of 0.07, which is much closer to the

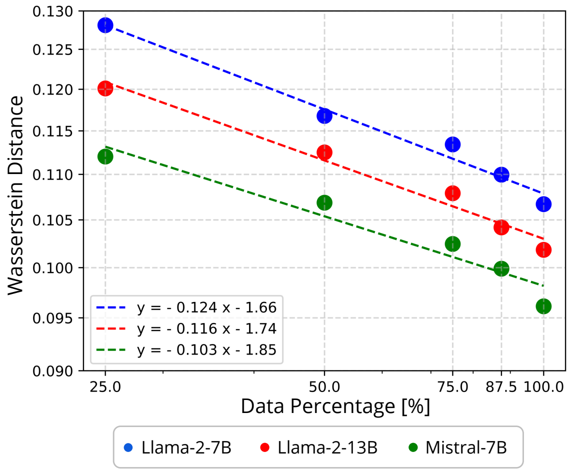

Figure 4: Evaluation results on OpinionQA after fine-tuning each LLM on increasingly large sampled subsets of SubPOP-Train. The plot x-axis is the size of sampled dataset and y-axis is WD against human responses measured on OpinionQA. Note that both axes are log scale. Dashed lines represent a line of best fit. Performances at data percentage of 100% are identical to ours in Table 1.

empirical lower bound of 0.031 reported in Table 1. This result underscores the critical importance of collecting more high-quality data, as increased dataset size can drive significant improvements in model performance.

5 Conclusion

In this work, we demonstrated that fine-tuning large language models on structured public opinion survey data markedly improves their ability to predict human response distributions. We curate SubPOP —a dataset 6.5× larger than previous collections to fine-tune and evaluate LLMs on survey response distribution prediction task. By training on SubPOP, we showed that LLMs can accurately capture the nuanced, group-specific variability in public opinions, while also generalizing to unseen survey waves and different survey families. Our experiments reveal that as the fine-tuning dataset grows, model performance continues to scale favorably, underscoring the importance of dataset size and representative sampling strategies.

These findings not only advance the state of opinion prediction but also highlight a broader societal imperative: to support public opinion research and survey design, there is a critical need to invest in and collect high-quality, large-scale survey data. Such efforts will enable more accurate modeling of diverse human opinions and, in turn, assist more informed decision-making in both public policy and research contexts.

6 Limitations

In this work, we explore the capability of language models to complement traditional survey design by predicting survey responses in advance. However, we acknowledge the following inherent limitations of this approach.

Role in Survey Research. While language models can provide a coarse approximation of human opinions, they cannot fully replace human involvement in the survey process. Human opinions evolve dynamically in response to social events, and while pretrained language models can incorporate such knowledge through retrieval-augmented generation, they remain limited in adapting to a rapidly changing world. Moreover, fine-tuning a language model on distributions of human opinions may inadvertently replicate and amplify existing biases of humans, leading to undesirable outcomes. It is important to note that a model fine-tuned on human opinions does not necessarily align with human values and behaviors, nor does it serve as a perfect proxy for human decision-making. The scope of our work is restricted to language models prompted with a group-level information generating response distributions to survey questions, rather than simulating individual human respondents in a personalized manner.

Data Dependence. Survey response data, even after post-stratification calibration, remain subject to empirical variance, particularly for relatively small groups that comprise about one percent of the U.S. population. Also, while traditional surveys have implemented various strategies to mitigate response bias stemming from the linguistic and multiple-choice nature of survey questions (Tourangeau, 2000), the extent to which these biases affect language models—and how best to address them—remains an open question (Tjuatja et al., 2024; Bisbee et al., 2024). Future research could focus on developing reliable opinion datasets for underrepresented groups and examining how prompt engineering elements can be optimized to reduce bias in language model-generated responses.

Limited Contextual Information. Our fine-tuning approach, which structures prompts in a `QA` format, demonstrates strong matching with human opinion distributions. However, we have not explored fine-tuning with richer contextual information. Prior research suggests that incorporating additional contextual details can improve the fidelity of model-generated opinions to actual human responses. We anticipate that more sophisticated steering techniques could further enhance the opinion prediction performance beyond the results presented in this study. Investigating such methods remains an open and promising direction for future work.

7 Potential Risks

Employing language models for opinion prediction has both influential possibilities and risk of misuse. We acknowledge that the risk of misuse cannot be overlooked, and we clearly state that indiscriminately minimizing the discrepancy of opinion response distribution as a fine-tuning target can cause severe harms. In particular, the model might develop a bias toward specific demographics during the course of fine-tuning, an artifact of minimizing response distribution when other safeguard measures are not employed. We emphasize that an oversight and holistic evaluation of methods and pipelines are required before deploying such models for any of the actual applications and interactions with human.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Sotiris Anagnostidis and Jannis Bulian. 2024. How susceptible are llms to influence in prompts? *arXiv preprint arXiv:2408.11865*.

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

Ashwini Ashokkumar, Luke Hewitt, Isaias Ghezae, and Robb Willer. 2024. Predicting results of social science experiments using large language models. *accessed September*, 19:2024.

Christopher A Bail. 2024. Can generative ai improve social science? *Proceedings of the National Academy of Sciences*, 121(21):e2314021121.

Jelke Bethlehem. 2010. Selection bias in web surveys. *International statistical review*, 78(2):161–188.

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

Pew Research Center. 2018. America trends panel waves. Retrieved February 06, 2025, from https://www.pewsocialtrends.org/dataset.

Souradip Chakraborty, Jiahao Qiu, Hui Yuan, Alec Koppel, Dinesh Manocha, Furong Huang, Amrit Bedi, and Mengdi Wang. 2024. Maxmin-rlhf: Alignment with diverse human preferences. In *Forty-first International Conference on Machine Learning*.

Daiwei Chen, Yi Chen, Aniket Rege, and Ramya Korlakai Vinayak. 2024. Pal: Pluralistic alignment framework for learning from heterogeneous preferences. *arXiv preprint arXiv:2406.08469*.

Myra Cheng, Esin Durmus, and Dan Jurafsky. 2023a. Marked personas: Using natural language prompts to measure stereotypes in language models. *arXiv preprint arXiv:2305.18189*.

Myra Cheng, Tiziano Piccardi, and Diyi Yang. 2023b. Compost: Characterizing and evaluating caricature in llm simulations. *arXiv preprint arXiv:2310.11501*.

Cheng-Han Chiang and Hung-yi Lee. 2023. Can large language models be an alternative to human evaluations? *arXiv preprint arXiv:2305.01937*.

Bernard CK Choi and Anita WP Pak. 2004. A catalog of biases in questionnaires. *Preventing chronic disease*, 2(1):A13.

Eric Chu, Jacob Andreas, Stephen Ansolabehere, and Deb Roy. 2023. Language models trained on media diets can predict public opinion. *arXiv preprint arXiv:2303.16779*.

Michael Davern, Rene Bautista, Jeremy Freese, Pamela Herd, and Stephen L. Morgan. 2024. General social survey 1972-2024. Principal Investigator: Michael Davern; Co-Principal Investigators: Rene Bautista, Jeremy Freese, Pamela Herd, and Stephen L. Morgan. Sponsored by National Science Foundation. NORC ed. Chicago: NORC, 2024.

Ameet Deshpande, Vishvak Murahari, Tanmay Rajpurohit, Ashwin Kalyan, and Karthik Narasimhan. 2023. Toxicity in chatgpt: Analyzing persona-assigned language models. *arXiv preprint arXiv:2304.05335*.

Danica Dillion, Niket Tandon, Yuling Gu, and Kurt Gray. 2023. Can ai language models replace human participants? *Trends in Cognitive Sciences*, 27(7):597–600.

Ricardo Dominguez-Olmedo, Moritz Hardt, and Celestine Mendler-Dünner. 2023. Questioning the survey responses of large language models. *arXiv preprint arXiv:2306.07951*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Esin Durmus, Karina Nyugen, Thomas I Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, et al. 2023. Towards measuring the representation of subjective global opinions in language models. *arXiv preprint arXiv:2306.16388*.

Shangbin Feng, Chan Young Park, Yuhan Liu, and Yulia Tsvetkov. 2023. From pretraining data to language models to downstream tasks: Tracking the trails of political biases leading to unfair nlp models. *arXiv preprint arXiv:2305.08283*.

Shangbin Feng, Taylor Sorensen, Yuhan Liu, Jillian Fisher, Chan Young Park, Yejin Choi, and Yulia Tsvetkov. 2024. Modular pluralism: Pluralistic alignment via multi-LLM collaboration. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4151–4171, Miami, Florida, USA. Association for Computational Linguistics.

Isabel O Gallegos, Ryan A Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K Ahmed. 2024. Bias and fairness in large language models: A survey. *Computational Linguistics*, pages 1–79.

Igor Grossmann, Matthew Feinberg, Dawn C Parker, Nicholas A Christakis, Philip E Tetlock, and William A Cunningham. 2023. Ai and the transformation of social science research. *Science*, 380(6650):1108–1109.

Perttu Hämäläinen, Mikke Tavast, and Anton Kunnari. 2023. Evaluating large language models in generating synthetic hci research data: a case study. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–19.

Zihao He, Minh Duc Chu, Rebecca Dorn, Siyi Guo, and Kristina Lerman. 2024. Community-cross-instruct: Unsupervised instruction generation for aligning large language models to online communities. *arXiv preprint arXiv:2406.12074*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

EunJeong Hwang, Bodhisattwa Prasad Majumder, and Niket Tandon. 2023. Aligning language models to user opinions. *arXiv preprint arXiv:2305.14929*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Liwei Jiang, Taylor Sorensen, Sydney Levine, and Yejin Choi. 2024. Can language models reason about individualistic human values and preferences? *arXiv preprint arXiv:2410.03868*.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.

Shivani Kapania, William Agnew, Motahhare Eslami, Hoda Heidari, and Sarah Fox. 2024. 'simulacrum of stories': Examining large language models as qualitative research participants. *arXiv preprint arXiv:2409.19430*.

Jaehyung Kim and Yiming Yang. 2024. Few-shot personalization of llms with mis-aligned responses. *arXiv preprint arXiv:2406.18678*.

Junsol Kim and Byungkyu Lee. 2023. Ai-augmented surveys: Leveraging large language models and surveys for opinion prediction. *arXiv preprint arXiv:2305.09620*.

Hannah Rose Kirk, Alexander Whitefield, Paul Röttger, Andrew Bean, Katerina Margatina, Juan Ciro, Rafael Mosquera, Max Bartolo, Adina Williams, He He, et al. 2024. The prism alignment project: What participatory, representative and individualised human feedback reveals about the subjective and multicultural alignment of large language models. *arXiv preprint arXiv:2404.16019*.

Thom Lake, Eunsol Choi, and Greg Durrett. 2024. From distributional to overton pluralism: Investigating large language model alignment. *arXiv preprint arXiv:2406.17692*.

J. Learner. 2024. The promise and pitfalls of AI-augmented survey research. Blog post. NORC at the University of Chicago. Retrieved from www.norc.org.

Cheng Li, Mengzhou Chen, Jindong Wang, Sunayana Sitaram, and Xing Xie. 2024. Culturellm: Incorporating cultural differences into large language models. *arXiv preprint arXiv:2402.10946*.

Junyi Li, Ninareh Mehrabi, Charith Peris, Palash Goyal, Kai-Wei Chang, Aram Galstyan, Richard Zemel, and Rahul Gupta. 2023. On the steerability of large language models toward data-driven personas. *arXiv preprint arXiv:2311.04978*.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Teaching models to express their uncertainty in words. *arXiv preprint arXiv:2205.14334*.

I Loshchilov. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Benjamin S Manning, Kehang Zhu, and John J Horton. 2024. Automated social science: Language models as scientist and subjects. Technical report, National Bureau of Economic Research.

Nicole Meister, Carlos Guestrin, and Tatsunori Hashimoto. 2024. Benchmarking distributional alignment of large language models. *arXiv preprint arXiv:2411.05403*.

Igor Melnyk, Youssef Mroueh, Brian Belgodere, Mattia Rigotti, Apoorva Nitsure, Mikhail Yurochkin, Kristjan Greenewald, Jiri Navratil, and Jerret Ross. 2024. Distributional preference alignment of llms via optimal transport. *arXiv preprint arXiv:2406.05882*.

Andrew Mercer, Arnold Lau, and Courtney Kennedy. 2018. For weighting online opt-in samples, what matters most?

Suhong Moon, Marwa Abdulhai, Minwoo Kang, Joseph Suh, Widyadewi Soedarmadji, Eran Kohen Behar, and David Chan. 2024. Virtual personas for language models via an anthology of backstories. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 19864–19897, Miami, Florida, USA. Association for Computational Linguistics.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Joon Sung Park, Carolyn Q Zou, Aaron Shaw, Benjamin Mako Hill, Carrie Cai, Meredith Ringel Morris, Robb Willer, Percy Liang, and Michael S Bernstein. 2024a. Generative agent simulations of 1,000 people. *arXiv preprint arXiv:2411.10109*.

Peter S Park, Philipp Schoenegger, and Chongyang Zhu. 2024b. Diminished diversity-of-thought in a standard large language model. *Behavior Research Methods*, pages 1–17.

Pew Research Center. 2024. Pew research center. Accessed: February 10, 2025.

Sriyash Poddar, Yanming Wan, Hamish Ivison, Abhishek Gupta, and Natasha Jaques. 2024. Personalizing reinforcement learning from human feedback with variational preference learning. *arXiv preprint arXiv:2408.10075*.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36.

Arun Rajkumar and Shivani Agarwal. 2014. A statistical convergence perspective of algorithms for rank aggregation from pairwise data. In *International conference on machine learning*, pages 118–126. PMLR.

David M. Rothschild, James Brand, Hope Schroeder, and Jenny Wang. 2024. Opportunities and risks of llms in survey research. Available on SSRN: http://dx.doi.org/10.2139/ssrn.5001645.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose opinions do language models reflect? In *International Conference on Machine Learning*, pages 29971–30004. PMLR.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. 2023. Quantifying language models' sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting. *arXiv preprint arXiv:2310.11324*.

Gabriel Simmons. 2022. Moral mimicry: Large language models produce moral rationalizations tailored to political identity. *arXiv preprint arXiv:2209.12106*.

Anand Siththaranjan, Cassidy Laidlaw, and Dylan Hadfield-Menell. 2023. Distributional preference learning: Understanding and accounting for hidden context in rlhf. *arXiv preprint arXiv:2312.08358*.

Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, et al. 2024. A roadmap to pluralistic alignment. *arXiv preprint arXiv:2402.05070*.

Chenkai Sun, Ke Yang, Revanth Gangi Reddy, Yi R Fung, Hou Pong Chan, Kevin Small, ChengXiang Zhai, and Heng Ji. 2024. Persona-db: Efficient large language model personalization for response prediction with collaborative data refinement. *arXiv preprint arXiv:2402.11060*.

Lindia Tjuatja, Valerie Chen, Tongshuang Wu, Ameet Talwalkwar, and Graham Neubig. 2024. Do llms exhibit human-like response biases? a case study in survey design. *Transactions of the Association for Computational Linguistics*, 12:1011–1026.

Roger Tourangeau. 2000. The psychology of survey response. *University of Cambridge*.

Angelina Wang, Jamie Morgenstern, and John P Dickerson. 2024. Large language models cannot replace human participants because they cannot portray identity groups. *arXiv preprint arXiv:2402.01908*.

World Values Survey. 2022. World Values Survey. [Online; accessed 02/15/2025].

Binwei Yao, Zefan Cai, Yun-Shiuan Chuang, Shanglin Yang, Ming Jiang, Diyi Yang, and Junjie Hu. 2024. No preference left behind: Group distributional preference optimization. *arXiv preprint arXiv:2412.20299*.

Siyan Zhao, John Dang, and Aditya Grover. 2023. Group preference optimization: Few-shot alignment of large language models. *arXiv preprint arXiv:2310.11523*.

Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2024. Can large language models transform computational social science? *Computational Linguistics*, 50(1):237–291.

Ai Assistants In Writing: We have used AI assistants (ChatGPT) in our writing.

A Dataset Details

A.1 American Trends Panel Datasets

Pew Research holds regular American Trends Panel (ATP) survey (called waves) (Center, 2018) covering various topics (*e.g.* veterans, political priorities, gender and leadership) and releases result at an individual level. For each anonymized individual, the following information is released: unique identification number, demographic details, survey responses, and weight. Weights (Mercer et al., 2018) are the output of post-survey calibration process that helps adjusting survey results for response bias (e.g., non-response bias, sampling bias) correction and population representativeness. As of January 2025, survey data until wave 132 has been released. About 20 surveys are conducted in each year.

A.2 OpinionQA

OpinionQA is a subset of ATP curated in (Santurkar et al., 2023). This dataset consists of contentious 500 questions sampled from 14 ATP waves which have high inter-group disagreement (i.e. large Wasserstein distances among demographic groups to a question). It also comes with hand-crafted ordinality information which provides structure to option lists. For example, options 'Major reason', 'Minor reason', and 'Not a reason', are assigned an ordinality mapping to 1, 2, and 3, respectively. This ordinality allows a calculation of 1-dimensional Wasserstein distance.

Demographic groups we employ and the number of questions per each of 14 waves are listed in Table 3. This set of groups are adopted for several small-scale analysis (Santurkar et al., 2023; Zhao et al., 2023; Kim and Yang, 2024). We note that our approach is not limited to a specific number of groups and data is available for minority or fine-grained demographic subpopulations.

A.3 SubPOP

We gather additional data from the American Trends Panel, specifically collecting 53 waves from Wave 61 to 132. There are 62 waves from Wave 61 - 132, however, some waves have missing demographic or ideology information (for example, wave 63 does not contain political ideology information) or the data is not available hence removed during the curation process. To refine the dataset, we exclude questions that meet the

Table 3: A list of 22 demographic groups and a wave-level information for waves included in OpinionQA dataset.

Trait	Groups	Population % in Wave 82
Region	Northeast	17.2
	South	37.8
Education	College grad+	24.2
	Less than high school	5.2
Gender	Male	44.3
	Female	54.6
Race / ethnicity	Black	9.6
	White	66.1
	Asian	4.8
	Hispanic	15.2
Income	$100,000 or more	21.8
	Less than $30,000	21.3
Political Party	Democrat	35.1
	Republican	29.1
Political Ideology	Liberal	20.0
	Conservative	22.6
	Moderate	38.3
Religion	Protestant	40.8
	Jewish	2.0
	Hindu	0.9
	Atheist	0.6
	Muslim	0.7

Wave	# questions	Wave Topic
26	44	Guns
29	20	Views on gender
32	24	Community types, Sexual harassment
34	16	Biomedical and food issues
36	68	Gender and leadership
41	41	Views of America in 2050
42	26	Trust in science
43	51	Race in America
45	13	Misinformation
49	19	Privacy and surveillance
50	43	American families
54	50	Economic inequality
82	56	2021 Global Attitudes Project U.S. survey
92	23	Political Typology

following criteria: those with more than 10 response options, redacted response data, or dependencies on prior questions (e.g., assessing political strength). For the remaining questions, we use GPT-4o to refine their wording, ensuring they are well-suited for individual prompting while making minimal modifications. In Figure 5 we provide a few-shot prompt for question refinement.

In Figure 6, we visualize the embeddings of the question texts (projected to 2-dimensions using t-SNE) from OpinionQA compared to the ATP and GSS portions of SubPOP. The visualization shows how much larger our dataset is than OpinionQA (6.5×), along with the expanded coverage of our dataset into semantic areas untouched by OpinionQA. The embeddings also reveal the distribution shift from ATP questions to GSS questions: while the ATP and GSS questions overlap in embedding space, the GSS question appear as small clusters, not evenly distributed over the ATP questions.

Instruction: Refine the question with a minimal change to make the question sensible. Do not modify options, and do not modify a question if it makes sense. Always start your answer with "Refined question:".

Question: A cross // Do you have any of the following for spiritual purposes?
A. Yes, I have this for spiritual purposes
B. No, I do not have this for spiritual purposes

Refined question: Do you have a cross for spiritual purposes?

Question: As you may know, same-sex marriage is now legal in the U.S. Do you think this is [a good thing or a bad thing] for our society?
A. Very good thing
B. Somewhat good thing
C. Somewhat bad thing
D. Very bad thing

Refined question: As you may know, same-sex marriage is now legal in the U.S. Do you think this is a good thing or a bad thing for our society?,

Question: On a different subject...How much, if at all, do white people benefit from advantages in society that black people do not have
A. A great deal
B. A fair amount
C. Not too much
D. Not at all

Refined question: How much, if at all, do white people benefit from advantages in society that black people do not have?,

Question: Thinking about the past couple of weeks, would you say the news for Donald Trump has been...
A. Very good
B. Mostly good
C. Neither good nor bad
D. Mostly bad
E. Very bad

Refined question: Thinking about the past couple of weeks, would you say the news for Donald Trump has been...

Question: **(Question to refine)**
(Options)

Refined question:

Figure 5: Few-shot prompt for refining the question to suit a language model prompting. An instruction is designed to make a minimal change to the original question, and in-context examples are provided.

A.4 General Social Survey 2022

To evaluate the out-of-distribution generalization ability of our fine-tuned models, we subsample 133 questions from the GSS 2022 dataset (Davern et al., 2024). We apply the same selection criteria as outlined in Appendix A.3, excluding questions that are redacted, conditioned on prior questions, directly answered through demographic steering, derived from a set of questions, or those with more than 10 response options.

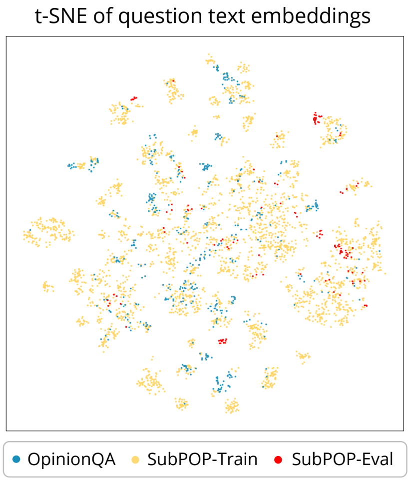

Figure 6: Embeddings of questions from OpinionQA, SubPOP-Train, and SubPOP-Eval.

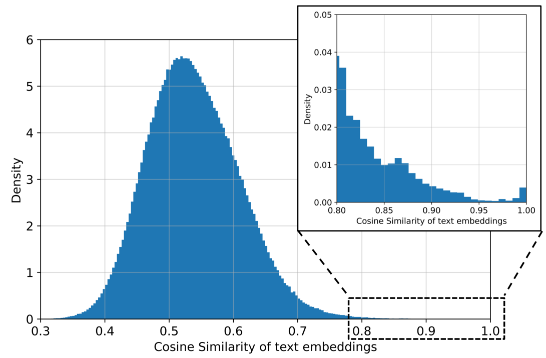

Figure 7: Distribution of cosine similarities between a question in SubPOP-ATP and OpinionQA, having a long tail towards a high cosine similarity. We inspect the question pairs in the range of 0.8 to 1.0 (distribution shown in the magnified view) and used a similarity of 0.87 as a safe threshold to identify a semantically identical question pair.

A.5 Inspection of Identical Questions

Distribution of cosine similarities between two text embeddings (an output of the embedding model OpenAI-text-embedding-3-large given a question), one from a question in SubPOP and another from a question in OpinionQA is shown in Figure 7. We observe a fraction of pairs having high cosine similarity, and manually inspected question pairs with high relevance pairs and find that by setting a threshold cosine similarity of 0.87 we can detect all semantically identical pairs. We took a conservative threshold of cosine similarity; this value was to maximize the recall at a cost of precision to ensure detection of overlapping questions.

B Training Details

We conduct our experiments using Nvidia A100 GPUs with 80GB VRAM. Hyperparameter tuning is performed over learning rates {5e-5, 1e-4, 2e-4} and batch sizes {64, 128, 256}. After evaluating possible combinations, we select a (learning rate, batch size) = (2e-4, 256) for Llama-2-7B, (learning rate, batch size) = (2e-4, 256) for Mistral-7B-v0.1, and (learning rate, batch size) = (1e-4, 256) for Llama-2-13B when utilizing the full training dataset. For Llama-3-70B, we have not done hyperparameter search but heuristically used (learning rate, batch size) = (2e-5, 256).

For sub-sampled training data (Figure 4), we use the following configurations:

- (lr, bs) = (2e-4, 256) for 75% of the training data
- (lr, bs) = (1e-4, 128) for 50% of the training data
- (lr, bs) = (1e-4, 128) for 25% of the training data

All training is performed using LoRA (Hu et al., 2021), with LoRA parameters initialized from a normal distribution with $\sigma = 0.02$. We set the LoRA rank to 8, alpha to 32, and apply a dropout rate of 0.05. LoRA weights are applied to the query and value matrices. The AdamW (Loshchilov, 2017) optimizer is used with a weight decay of 0.

B.1 Choice of the training objective

In this section, we explore both forward KL-divergence and Wasserstein Distance (WD) as training objectives. The forward KL-divergence is defined as

$$D_{\mathrm{KL}}(p_H \| p_\theta) = \sum_{a \in \mathcal{A}_q} p_H(a) \log \frac{p_H(a)}{p_\theta(a)},$$

where $p_H(a) \equiv p_H(a \mid q, g)$ and $p_\theta(a) \equiv p_\theta(a \mid q, g)$. Similarly, WD is given by

$$\mathcal{WD}(p_H, p_\theta) = \min_{\gamma \in \Pi(p_H, p_\theta)} \sum_{a, a' \in \mathcal{A}_q} \gamma(a, a') d(a, a'),$$

with $\Pi(p_H, p_\theta)$ denoting the set of all couplings between p_H and p_θ, and $d(a, a')$ the L1 distance between choices. Since survey responses are inherently one-dimensional and ordinal, we can simplify the computation of WD using cumulative distribution functions (CDFs). In the 1-D case, WD is computed as

$$\mathcal{WD}(p_H, p_\theta) = \int_{-\infty}^{+\infty} |F_{p_H}(x) - F_{p_\theta}(x)| dx,$$

$$= \sum_{i=1}^{n} |F_{p_H}(i) - F_{p_\theta}(i)|$$

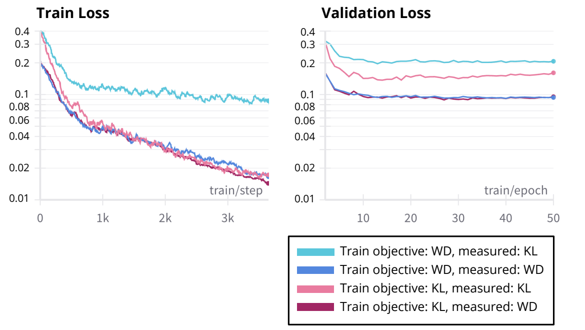

Figure 8: Train loss curve (left) and validation loss curve (right) for Llama-2-7B fine-tuned on 90% of OpinionQA, with the remaining 10% used for validation. Light and dark blue lines represent KL-divergence (KL) and Wasserstein distance (WD) when used KL as a training objective, while light and dark red lines represent KL and WD when used WD as a training objective. The two training objectives yield similar results in terms of WD, the primary measure of opinion distribution matching in our work.

where F_{p_H} and F_{p_θ} are the CDFs corresponding to p_H and p_θ, respectively. We use this discrete formulation as the WD loss in our training.

While training with WD resulted in a higher KL-divergence on the validation set, the validation WD converged to similar levels regardless of the objective (see Figure 8). We attribute this to KL-divergence penalizing low-probability assignments without significantly altering the overall distribution geometry. Given its broader applicability—without requiring ordinal information—we primarily used KL-divergence in our experiments.

However, the choice of objective did not significantly impact the opinion prediction performance, as measured by WD (Figure 8). Although WD as a training objective resulted in higher KL-divergence on the validation set, the validation WD converged to the same level regardless of the training objective. We attribute this to KL-divergence strongly penalizing language models' probability assignments to options with low human opinion probability (choices rarely selected by humans). However, since these probabilities remain low, the shape of probability distribution is preserved. Given KL-divergence's broader applicability—it does not require ordinal information—we primarily used it in our experiments.

C Additional Experiments

C.1 Effect of Response Distribution Modeling

In this section, we compare different methods for capturing the distribution of human responses. We consider three approaches:

1. *One-hot*: Predicting only the most probable response, which ignores the full distribution over all responses (Li et al., 2024).

2. *Augment by N*: Augmenting the dataset by replicating each response by a factor of N according to its observed frequency (Zhao et al., 2023).

3. *Explicit probability modeling*: Directly modeling the full response distribution using the actual probability values for each option.

Table 4 summarizes the results of these approaches. Notably, the explicit probability modeling outperforms one-hot with a considerable margin. This shows that merely learning the single most frequent response fails to capture the opinion diversity within each demographic subgroup.

Compared with augmented data, the explicit modeling performs better than the augmentation approach. Notably, the performance gap is larger than the quantization error introduced by discretizing the response distribution. If we use N for discretization, the quantization error is $\frac{1}{2N}$, which is continuous value with 0.01 or 0.005 for the cases in Table 4. Also, the other benefit of explicit modeling compared to augment by N, is that we can reduce the amount of data by a factor of N. This reduces the cost of fine-tuning LLMs.

Table 4 summarizes the results of these approaches. Notably, explicit probability modeling substantially outperforms the one-hot method, demonstrating that simply predicting the single most frequent response fails to capture the opinion diversity present within each demographic subgroup.

Compared with augment by N (2nd and 3rd column in Table 4), explicit probability modeling also achieves better performance. Importantly, the performance gap exceeds the quantization error introduced by discretizing the response distribution. For instance, when discretizing with a factor of N, the quantization error is $\frac{1}{2N}$—approximately 0.01 or 0.005 in the cases shown in Table 4. Moreover, explicit modeling offers the practical benefit of reducing the data volume by a factor of N compared to the augmentation approach, thereby lowering the computational cost of fine-tuning LLMs.

These results underscore the importance of explicit distribution modeling. By aligning the model's predictive distribution directly with the survey distribution, we achieve higher accuracy with fewer data samples, avoiding the rounding errors and replication overheads that are inherent to data-augmentation approaches.

C.2 Post-trained Model

We fine-tune Llama-2-7B-chat to observe the effect of starting from checkpoints that have been instruction-tuned via Reinforcement Learning from Human Feedback (RLHF). Table 5 shows the evaluation performance of a baseline method (Zero-shot prompting (QA)), fine-tuned base model and our fine-chat model. We observe the significant performance improvement, while the baseline method performs worse then the models not instruction-tuned (Table 1). Especially, the performance for SubPOP-Eval of chat model is significantly worse than that of base model. We observe the high WD of the baseline method resulting from the model assigning high probability to a specific token (e.g. 'A'), being far apart from the human opinion distribution. After fine-tuning the model are able to generate a more distributed probability of answer tokens. This result coincides with the result reported in (Moon et al., 2024).

C.3 Generalization to Unseen Subpopulations

Here we present a complete list of evaluation performance on OpinionQA for unseen subpopulations (the groups not used to fine-tune our model) and perform an analysis that shows our fine-tuned models are able to steer towards the given subpopulation information.

As shown in Table 6, we observe a performance improvement across unseen subpopulations. To verify that the performance improvement does not come from the model simply utilizing average opinion distribution (average of response distributions across subpopulations used in the fine-tuning data),

Table 4: Comparison of evaluation performance for three response distribution modeling approaches, with Llama-2-7B as a base model. The last column (Explicit) is identical to the ours presented in Table 1. A model fine-tuned to predict the most probable choice (one-hot) performs the worst, as the model has not learned distributional opinion at fine-tuning phase. A model trained on augmented data (Aug. ($\times$50, $\times$100)), while performing much better than one-hot still underperforms the explicit distribution modeling.

Eval Dataset	One-hot	Aug. ($\times$ 50)	Aug. ($\times$ 100)	Explicit (Ours)
OpinionQA	0.163	0.110	0.107	0.106
SubPOP-Eval	0.178	0.130	0.123	0.121

Table 5: Performance of the fine-tuned Llama-2-7B-chat model (Chat LLM). For comparison, we also present lower and upper bounds, the baseline method Zero-shot prompt (QA) and fine-tuned Llama-2-7B (Base LLM).

Method	**OpinionQA**	**SubPOP-Eval**
Upper bound (Unif.)	0.178	0.208
Lower bound (Human)	0.031	0.033
Zero-shot prompt (QA)	0.308	0.383
Chat LLM	0.109	0.148
Base LLM	0.106	0.121

we perform an analysis of how closely a fine-tuned model provided with a steering prompt for group X represents the response distribution for group Y.

To verify that the observed improvements are not coincidental, we analyze in Figure 9, Figure 10, Figure 11, and Figure 12 how well language models conditioned with different steering prompts match the true distributions of various subpopulations. Concretely, we measure how closely our fine-tuned model provided with a steering prompt of group X predict response distribution of human group Y. We observe that even for unseen subpopulations Y should be X to minimize the WD between the model's response distribution and the response of group X, confirming that the model tailors its predictions to each unseen group rather than defaulting to an averaged distribution of . We hypothesize that this is possible because the model learns to be jointly conditioned on the subpopulation information and survey question during fine-tuning, and also utilizing its knowledge on relationship between subpopulations, able to predict the opinion distribution even for unseen groups.

D Baseline Details

- **Upper bound**: We estimate the distribution between human responses and uniform distribution as an upper bound of WD metrics.
- **Zero-shot prompting**: Three prompt styles—QA, BIO, and PORTRAY—are introduced in (Santurkar et al., 2023) to integrate group information into prompts. These prompts are then combined with survey questions to construct inputs for LLM. Then, the first-token log-probability from LLM is measured to calculate the model's response distribution over options. In our baseline (and also in fine-tuning experiments) we focus on the QA steering format. Examples of this prompting method are shown in Figure 13.
- **Few-shot prompting**: We craft a conditioning

Table 6: Evaluation performance on OpinionQA with demographics not included in the fine-tuning dataset SubPOP-training from Llama-2-7B. For reference, we present a lower bound (human) and the zero-shot prompting (QA) are presented. Absolute difference refers to the difference between zero-shot prompting and ours, and the relative improvement is caluclated in a same way to Figure 3.

Attribute	Group	Lower Bound (Human)	Zero-shot (QA)	Ours	Absolute Diff.	Relative Improvement
Age	18-29	0.023	0.185	0.096	0.089	0.548
Age	30-49	0.014	0.151	0.093	0.058	0.424
Age	50-64	0.014	0.154	0.101	0.052	0.377
Age	65+	0.013	0.195	0.115	0.080	0.438
Region	Midwest	0.016	0.153	0.095	0.058	0.425
Region	West	0.017	0.162	0.095	0.068	0.465
Education	Associate's Degree	0.026	0.159	0.098	0.061	0.455
Education	High School Graduate	0.017	0.144	0.092	0.053	0.413
Education	Postgraduate	0.015	0.174	0.106	0.068	0.426
Education	Some College, No Degree	0.018	0.144	0.093	0.051	0.405
Income	$50,000-$75,000	0.016	0.153	0.098	0.054	0.396
Income	$30,000-$50,000	0.019	0.144	0.094	0.050	0.400
Political Ideology	Very Conservative	0.026	0.208	0.107	0.101	0.555
Political Ideology	Very Liberal	0.025	0.202	0.111	0.091	0.514
Political Party	Independent	0.016	0.155	0.093	0.062	0.445
Political Party	Something Else	0.026	0.162	0.092	0.069	0.510
Race	Other	0.050	0.180	0.144	0.036	0.275
Religion	Agnostic	0.028	0.189	0.115	0.074	0.459
Religion	Buddhist	0.063	0.207	0.149	0.059	0.405
Religion	Nothing in Particular	0.019	0.153	0.092	0.061	0.454
Religion	Orthodox	0.083	0.221	0.180	0.041	0.298
Religion	Other	0.051	0.184	0.123	0.061	0.457
Religion	Roman Catholic	0.018	0.145	0.098	0.047	0.371

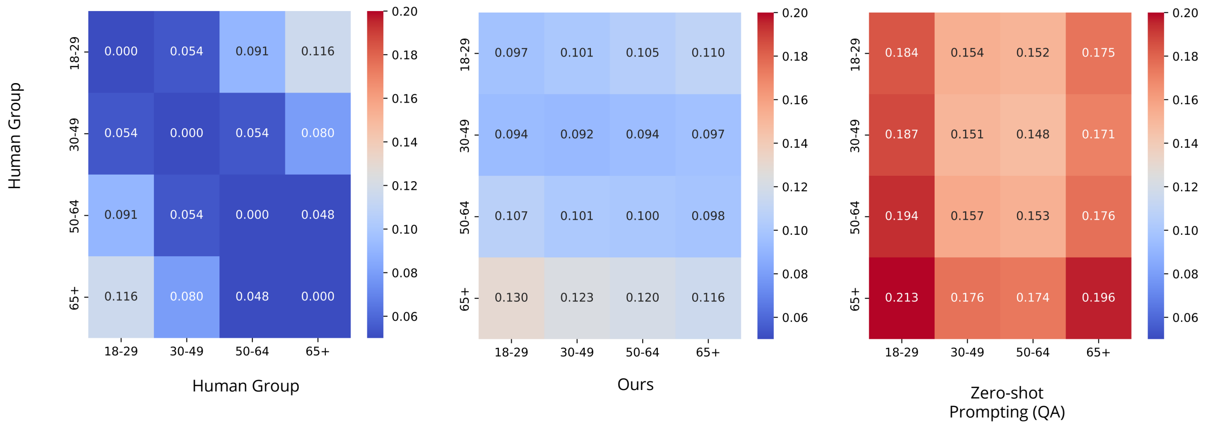

Figure 9: Heatmap of average WD between a human (y-axis) and a group on the x-axis for age trait. Our model, when steered with the conditioning prompt, exhibits similar WD pattern as between human groups, showing that our model are steered towards demographic subgroups.

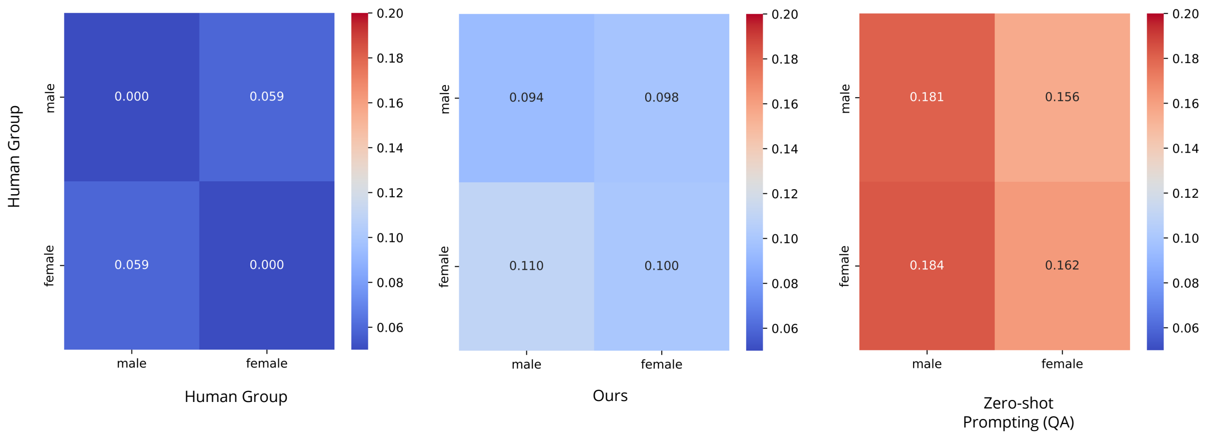

Figure 10: Heatmap of average WD between a human group (y-axis) and a group on the x-axis for gender trait.

prompt that contains not only group information but also the group's response distribution to k train questions, following (Hwang et al., 2023). For a test question $q_{test} \in Q_{test}$, we first sort training questions Q_{train} into $\{q_1, q_2, \ldots\}$ such that $\mathtt{sim}(\mathtt{E}(q_1), \mathtt{E}(q_{test})) > \mathtt{sim}(\mathtt{E}(q_2), \mathtt{E}(q_{test}))$, and so on. $\mathtt{E}(q)$ denotes the embedding model (OpenAI-text-embedding-3-large) output of the

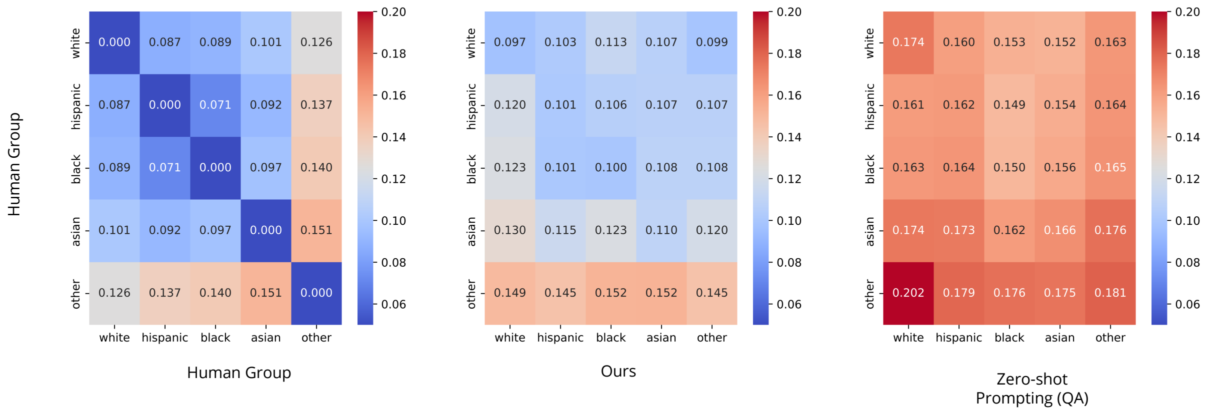

Figure 11: The heatmap of average WD between a human group (y-axis) and a group on the x-axis for race trait.

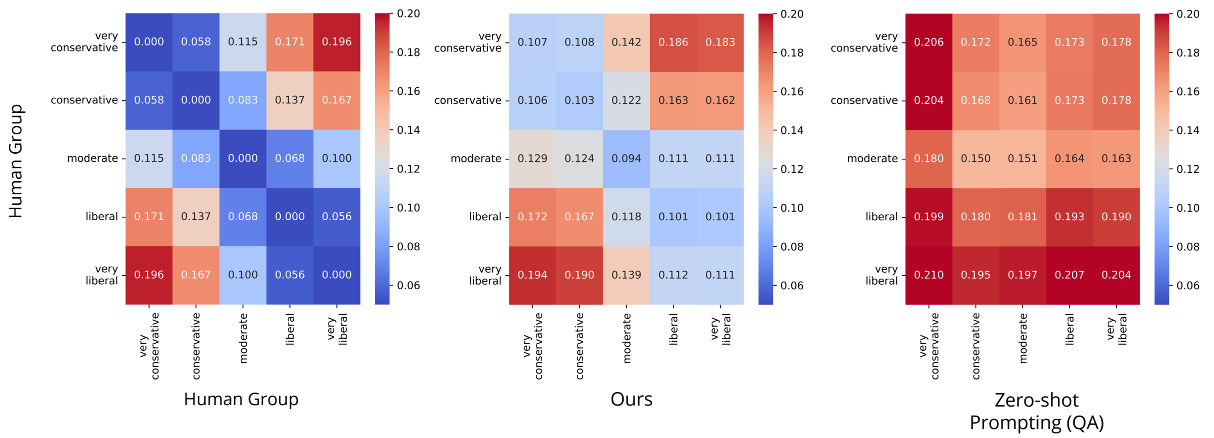

Figure 12: The heatmap of average WD between a human group (y-axis) and a group on the x-axis for political ideology trait.

input q and `sim` is a cosine similarity between two embedding vectors. Then, response information of the first k questions $\{q_i, p(\mathcal{A}_{q_i}|q_i, g)\}_{i=1}^{k}$ are used as few shot prompts to have the language model verbalize (Meister et al., 2024) expected response distribution for the given g and q_{test}. An example of the prompt for $k=3$ case is shown in Figure 14, while we run the baseline experiment in a $k=5$ setting.

- **Modular Pluralism**: The intuition behind Modular Pluralism (Feng et al., 2024) is that a language model trained on a text corpus of a specific sub-population will faithfully represent public opinion of that population. Given a survey question with a PORTRAY-style steering prompt, each of language model 'modules' (fine-tuned Mistral-7B-Instruct-v0.1) generates an option choice with explanation. A black-box LLM (GPT-3.5-turbo-Instruct) receives all generations and select a generation that best aligns with the given group. Finally, using the chosen generation as a context, a black-box LLM generates probability distribution over options. The example pipeline is shown in Figure 15. Instead of the sub-sampled OpinionQA dataset the authors of the method used, we use the exactly same evaluation set across all baseline methods and our approach for a fair comparison.

- **Lower bound**: We compute a lower bound by randomly sampling two groups from the human respondents and calculating the WD between their response distributions. Bootstrapping is then applied to obtain a robust estimate. Further details on this estimation process are provided below:

Computing weighted answer distributions: For each demographic group g and question q, we have n_{gq} responses from respondents who belong to group g answering question q: $x_1, x_2, \cdots, x_{n_{gq}}$, where $x_i \in \mathcal{A}_q$, i.e., the answer set for question q (e.g., $\{1,2,3,4\}$). Furthermore, each respondent (and thus, their response) is associated with a wave-specific weight $w_1, w_2, \cdots, w_{n_{gq}}$, provided by Pew Research. We compute the human answer distribution $\pi_{gq}^{(H)}$ as a weighted sum over responses, where the proportion of respondents providing answer $a \in \mathcal{A}_q$ is estimated as

$$\pi_{gq}^{(H)}(a) = \frac{\sum_{i=1}^{n_{gq}} w_i \mathbb{1}[x_i = a]}{\sum_{i=1}^{n_{gq}} w_i}.$$

Bootstrapping at the respondent-level: We draw bootstrap samples per demographic group at the respondent-level including questions from all survey waves. This allows us to capture correlations in answer distributions across questions and across waves.

Specifically, let $\mathcal{P}_g$ represent the set of respondents in group g, where $|\mathcal{P}_g| = n_g$. We produce bootstrapped samples by repeatedly sampling n_g respondents from $\mathcal{P}_g$ with replacement. Let $p_1^{(r)}, p_2^{(r)}, \cdots, p_{n_g}^{(r)}$ represent the sampled respondents for the r-th bootstrap, and let $w_1^{(r)}, w_2^{(r)}, \cdots, w_{n_g}^{(r)}$ represent their corresponding weights.

For each question q, let $\mathcal{P}_{gq} \subseteq \mathcal{P}_g$ represent the set of respondents from group g who answered question q; as before, $|\mathcal{P}_{gq}| = n_{gq}$. Let us define $q(p_i)$ as person p_i's response to question q if p_i answered question q, i.e., $p_i \in \mathcal{P}_{gq}$, and 0 otherwise. Then, we compute the r-th answer distribution for option a to question q as:

$$\pi_{gq}^{(r)}(a) = \frac{\sum_{i=1}^{n_g} \mathbb{1}[p_i^{(r)} \in \mathcal{P}_{gq}] w_i^{(r)} \mathbb{1}[q(p_i^{(r)}) = a]}{\sum_{i=1}^{n_g} \mathbb{1}[p_i^{(r)} \in \mathcal{P}_{gq}] w_i^{(r)}}.$$

Human lower bound of WD. Our statistic of interest is the mean Wasserstein distance over all questions Q across all waves per demographic group. We approximate this as the WD between the observed human distribution $\pi_{gq}^{(H)}$ and the bootstrap sample $\pi_{gq}^{(r)}$ for question q and group g. Over all R bootstraps, we have

$$\mathcal{D}_g^{(H)} = \left\{ \frac{1}{|Q|} \sum_{q \in Q} WD(\pi_{gq}^{(H)}, \pi_{gq}^{(r)}) \right\}_{r=1}^{R}.$$

To quantify agreement between human samples, we report the mean and 95% CI of $\mathcal{D}_{gq}^{(H)}$.

E Wave, Group-level Opinion Matching

Here we present a group-level and wave-level averaged Wasserstein distance. Wave-level result is in Table 9, and group-level results for OpinionQA and SubPOP-Eval are in Table 7, 8, respectively.

Question: Which part of the United States do you currently live in?
A. Northeast
B. Midwest
C. South
D. West

Answer: A. Northeast

Answer the following question keeping in mind your previous answers.
Question: I am going to name some institutions in this country. As far as the people running these institutions are concerned, would you say you have a great deal of confidence, only some confidence, or hardly any confidence at all in them? Military
A. A great deal
B. Only some
C. Hardly any
D. Refused
Answer as a choice between A.,B.,C.,D.

Answer:

Question: What is your present religion, if any?
A. Protestant
B. Roman Catholic
C. Mormon
D. Orthodox
E. Jewish
F. Muslim
G. Buddhist
H. Hindu
I. Atheist
J. Agnostic
K. Other
L. Nothing in particular

Answer: I. Atheist

Answer the following question keeping in mind your previous answers.
Question: Do you believe there is a life after death?
A. Yes
B. No
C. Refused
Answer as a choice between A.,B.,C.

Answer:

Figure 13: Two examples of LM steering in the QA format (Santurkar et al., 2023). Demographic group information (colored in pink) is concatenated with survey question (colored in sky blue). The first-token log-probability (probabilities assigned to A, B, C, ...) are used to calculate language model's response distribution. The same group steering prompt is used in our fine-tuning experiment.

Table 7: Per-group Wasserstein distance on OpinionQA for each base models, before and after fine-tuning on `SubPOP`-Train. Base refers to zero-shot prompting (`QA`).

Attribute	Group	Human Baseline	Llama-2-7B Base	Llama-2-7B Fine-tuned	Llama-2-13B Base	Llama-2-13B Fine-tuned	Mistral-7B-v0.1 Base	Mistral-7B-v0.1 Fine-tuned	Llama-3-70B Base	Llama-3-70B Fine-tuned
Region	Northeast	0.023	0.165	0.094	0.155	0.088	0.155	0.083	0.134	0.084
	South	0.017	0.149	0.092	0.143	0.085	0.133	0.081	0.113	0.078
Education	College grad, some Postgrad	0.018	0.165	0.099	0.157	0.096	0.136	0.089	0.125	0.085
	Less than high school	0.043	0.161	0.101	0.150	0.096	0.134	0.094	0.151	0.091
Gender	Male	0.015	0.182	0.093	0.152	0.089	0.131	0.083	0.138	0.083
	Female	0.013	0.162	0.100	0.158	0.092	0.146	0.088	0.130	0.087
Race / ethnicity	Black	0.031	0.151	0.102	0.144	0.095	0.132	0.091	0.116	0.085
	White	0.012	0.176	0.097	0.178	0.093	0.145	0.085	0.131	0.084
	Asian	0.051	0.165	0.111	0.167	0.104	0.143	0.102	0.124	0.099
	Hispanic	0.044	0.162	0.102	0.163	0.098	0.134	0.092	0.126	0.091
Income	$100,000 or more	0.019	0.172	0.103	0.162	0.100	0.147	0.091	0.159	0.087
	Less than $30,000	0.021	0.162	0.091	0.148	0.083	0.127	0.080	0.154	0.078
Political Party	Democrat	0.016	0.172	0.099	0.158	0.092	0.161	0.082	0.118	0.079
	Republican	0.019	0.196	0.105	0.235	0.101	0.181	0.095	0.174	0.093
Political Ideology	Liberal	0.022	0.192	0.100	0.181	0.094	0.166	0.084	0.126	0.081
	Conservative	0.021	0.169	0.103	0.153	0.099	0.144	0.094	0.141	0.092
	Moderate	0.016	0.151	0.094	0.153	0.090	0.132	0.082	0.106	0.081
Religion	Protestant	0.016	0.015	0.166	0.096	0.158	0.092	0.146	0.086	0.143
	Jewish	0.058	0.182	0.124	0.182	0.122	0.165	0.115	0.144	0.115
	Hindu	0.079	0.211	0.160	0.232	0.163	0.211	0.161	0.181	0.157
	Atheist	0.035	0.202	0.118	0.204	0.110	0.196	0.099	0.135	0.098
	Muslim	0.089	0.202	0.159	0.209	0.156	0.204	0.146	0.171	0.144

Table 8: Per-group Wasserstein distance on `SubPOP`-Eval for each base models, before and after fine-tuning on `SubPOP`-Train. Base refers to zero-shot prompting (`QA`).

Attribute	Group	Human Baseline	Llama-2-7B Base	Llama-2-7B Fine-tuned	Llama-2-13B Base	Llama-2-13B Fine-tuned	Mistral-7B-v0.1 Base	Mistral-7B-v0.1 Fine-tuned	Llama-3-70B Base	Llama-3-70B Fine-tuned
Region	Northeast	0.027	0.196	0.113	0.193	0.103	0.185	0.108	0.156	0.078
	South	0.018	0.183	0.108	0.185	0.103	0.176	0.103	0.138	0.080
Education	College grad, some Postgrad	0.019	0.206	0.105	0.175	0.101	0.167	0.099	0.137	0.077
	Less than high school	0.036	0.191	0.129	0.182	0.117	0.172	0.121	0.180	0.108
Gender	Male	0.017	0.186	0.102	0.176	0.101	0.170	0.099	0.150	0.079
	Female	0.016	0.184	0.108	0.198	0.105	0.176	0.100	0.151	0.080
Race / ethnicity	Black	0.029	0.200	0.114	0.179	0.102	0.170	0.107	0.139	0.094
	White	0.014	0.190	0.105	0.187	0.103	0.181	0.102	0.153	0.083
	Asian	0.049	0.201	0.119	0.190	0.107	0.184	0.114	0.158	0.096
	Hispanic	0.050	0.204	0.133	0.199	0.122	0.182	0.134	0.172	0.115
Income	$100,000 or more	0.021	0.210	0.111	0.184	0.106	0.176	0.102	0.179	0.082
	Less than $30,000	0.026	0.179	0.115	0.172	0.103	0.165	0.105	0.171	0.086
Political Party	Democrat	0.020	0.219	0.103	0.197	0.092	0.199	0.091	0.128	0.076
	Republican	0.023	0.205	0.123	0.234	0.117	0.206	0.115	0.187	0.093
Political Ideology	Liberal	0.019	0.224	0.102	0.191	0.090	0.188	0.096	0.134	0.076
	Conservative	0.022	0.184	0.120	0.178	0.112	0.172	0.113	0.160	0.092
	Moderate	0.018	0.191	0.110	0.183	0.103	0.170	0.103	0.141	0.082
Religion	Protestant	0.019	0.187	0.110	0.179	0.107	0.172	0.105	0.164	0.082
	Jewish	0.066	0.245	0.149	0.226	0.144	0.218	0.129	0.164	0.119
	Hindu	0.095	0.264	0.180	0.253	0.169	0.252	0.186	0.223	0.166
	Atheist	0.021	0.222	0.126	0.207	0.103	0.199	0.116	0.132	0.106
	Muslim	0.090	0.253	0.175	0.240	0.181	0.238	0.173	0.203	0.158

Table 9: Per-wave Wasserstein distance on OpinionQA for each base model, before and after fine-tuning on SubPOP-Train. Base refers to zero-shot prompting (QA).

Wave	Llama-2-7B		Llama-2-13B		Mistral-7B-v0.1		Llama-3-70B	
	Base	Fine-tuned	Base	Fine-tuned	Base	Fine-tuned	Base	Fine-tuned
26	0.191	0.145	0.180	0.126	0.178	0.131	0.134	0.084
29	0.169	0.096	0.172	0.123	0.153	0.096	0.125	0.085
32	0.163	0.110	0.156	0.098	0.137	0.099	0.151	0.091
34	0.155	0.105	0.171	0.089	0.134	0.095	0.138	0.083
36	0.175	0.120	0.184	0.126	0.175	0.107	0.130	0.087
41	0.160	0.090	0.155	0.084	0.134	0.073	0.116	0.085
42	0.159	0.053	0.146	0.059	0.127	0.059	0.131	0.084
43	0.179	0.112	0.172	0.104	0.154	0.102	0.124	0.099
45	0.177	0.101	0.177	0.093	0.149	0.084	0.126	0.091
49	0.151	0.098	0.143	0.131	0.128	0.116	0.159	0.087
50	0.209	0.139	0.196	0.121	0.188	0.125	0.154	0.078
54	0.158	0.087	0.158	0.087	0.128	0.077	0.118	0.079
82	0.173	0.098	0.171	0.075	0.148	0.077	0.174	0.093
92	0.165	0.073	0.153	0.071	0.140	0.055	0.126	0.081

Predict the answer distribution of the following questions for people whose **race or ethnicity** is **White**.

Question: Overall, how does being Asian affect people's ability to get ahead in our country these days?
A. Helps a lot
B. Helps a little
C. Hurts a little
D. Hurts a lot
E. Neither helps nor hurts
F. Refused

Answer distribution: {'A': 0.050, 'B': 0.210, 'C': 0.241, 'D': 0.034, 'E': 0.474, 'F': 0.010}

Question: Thinking about the people who design artificial intelligence computer programs, how well do you think they take into account the experiences and views of Asian adults?
A. Very well
B. Somewhat well
C. Not too well
D. Not at all well
E. Not sure
F. Refused

Answer distribution: {'A': 0.078, 'B': 0.216, 'C': 0.138, 'D': 0.069, 'E': 0.507, 'F': 0.008}

Question: Thinking about the people who design artificial intelligence computer programs, how well do you think they take into account the experiences and views of Hispanic adults?
A. Very well
B. Somewhat well
C. Not too well
D. Not at all well
E. Not sure
F. Refused

Answer distribution: {'A': 0.050, 'B': 0.160, 'C': 0.203, 'D': 0.090, 'E': 0.505, 'F': 0.008}

Question: Thinking again about race and race relations in the U.S. in general, how well, if at all, do you think each of these groups get along with each other in our society these days? Hispanics and Asians
A. Very well
B. Pretty well
C. Not too well
D. Not at all well
E. Not sure
F. Refused

Answer distribution:

Figure 14: Few-shot prompting example for $k = 3$. Demographic group information is presented in the beginning of the prompt (colored in pink). Following group information, k questions whose text embedding are the most similar to the text embedding of the evaluation question (colored in sky blue) are presented along with their opinion distribution. k questions are presented in the ascending order of cosine similarity. The generation of language model (verbalization of opinion distribution) is parsed to obtain the response distribution.

In terms of **political party**, you are **Republican**. Please respond to the following question with explanation.

Question: Please indicate whether the following is a major reason, a minor reason, or not a reason why you own a gun. As part of a gun collection
A. Major reason
B. Minor reason
C. Not a reason

Answer:

Which of the following comments best reflect the people of **Republican** in terms of **political party**?

Comment 1: **(generation from model 1)**

Comment 2: **(generation from model 2)**

Comment 3: **(generation from model 3)**

Comment 4: **(generation from model 4)**

Comment 5: **(generation from model 5)**

Comment 6: **(generation from model 6)**

Please select one comment number from 1 to 6.

In terms of **political party**, you are **Republican**. Please respond to the following question with the help of a passage.

Passage: **(selected generation)**

Question: Please indicate whether the following is a major reason, a minor reason, or not a reason why you own a gun. As part of a gun collection
A. Major reason
B. Minor reason
C. Not a reason

Answer:

Figure 15: Pipeline example of Modular Pluralism. Given a demographic group and a survey question, the first prompt is asked to multiple (6) language models, Mistral-7B-v0.1-Instruct fine-tuned on the community text corpus. The generations are sent to a black-box LLM (gpt-3.5-0613-Instruct) in the format of the second prompt. The black-box LLM answers which one of generations best reflects the given demographics. Finally, the selected generation serves as a context to answer the given survey question and the black-box LLM is prompted (the third prompt) to generate response distribution over the answer token A, B, C, etc.