Beyond Marginal Distributions: A Framework to Evaluate the Representativeness of Demographic-Aligned LLMs

Tristan Williams[1] **Franziska Weeber**[2] **Sebastian Padó**[2] **Alan Akbik**[1]

[1]Humboldt University of Berlin [2]University of Stuttgart
tdw75@outlook.com, {franziska.weeber|pado}@ims.uni-stuttgart.de, alan.akbik@hu-berlin.de

Abstract

Large language models are increasingly used to represent human opinions, values, or beliefs, and their steerability towards these ideals is an active area of research. Existing work focuses predominantly on aligning marginal response distributions, treating each alignment evaluation example independently. While essential, this may overlook deeper latent structures that characterise real populations and underpin cultural values theories. We propose a framework for evaluating the *representativeness* of aligned models through multivariate correlation patterns in addition to marginal distributions. We show the value of our evaluation scheme by comparing two model steering techniques (persona prompting and demographic fine-tuning) and evaluating them against human responses from the World Values Survey. While the demographic fine-tuned model better approximates marginal response distributions, persona prompting performs marginally better at reproducing the empirical correlation structure between survey items. Despite this reversal, neither technique aligns with human correlation patterns. We conclude that representativeness is a distinct aspect of value alignment and an evaluation focused on marginals can mask structural failures, leading to overly optimistic conclusions about model representativeness.

1 Introduction

As Large Language Models (LLMs) are increasingly integrated into socially sensitive domains, the challenge of pluralistic alignment, namely ensuring AI reflects the diverse intentions, ethical norms, and beliefs of a heterogeneous global population, has moved to the forefront of research (Gabriel, 2020; Weidinger et al., 2022; Sorensen et al., 2024; Lu and Kleek, 2024). A central difficulty for alignment lies in the non-monolithic nature of human values (Gabriel, 2020). Attempts to align AI systems with a monolithic conception of human values risk marginalising minority perspectives rather than preserving their inherent heterogeneity.

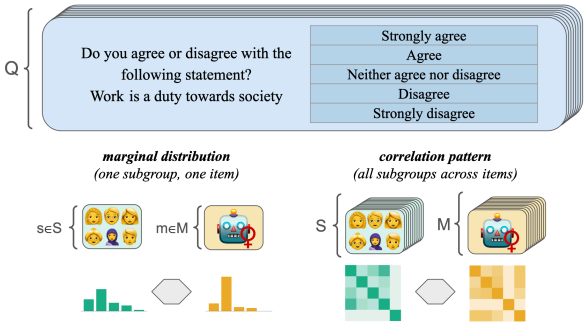

Figure 1: Overview of our suggested framework. Given a set of survey questions Q (top), we compare marginal responses of one human subgroup $s \in S$ and a steered model $m \in M$ for each question (left) as well as the correlation structures across subgroups S and across models M over all questions (right).

This paper engages directly with this problem by presenting a new methodology for evaluating *representativeness* using the popular medium of value surveys (Argyle et al., 2023; Santurkar et al., 2023; Bisbee et al., 2024; Durmus et al., 2024; Suh et al., 2025). We argue that the current focus in the literature on marginal distributions, i.e., comparing the response distributions of isolated questions, is a low bar for success. A truly representative model must capture not only what people think on each question individually, but also how their opinions are structured across questions. As noted in other fields, such as psycholinguistics (Hollis et al., 2017), focusing on isolated averages can mask significant structural failures. For example, a model might correctly approximate the support for two different policies independently, but fail to capture the fact that, in real populations, support for one is highly correlated with opposition to the other. Figure 1 represents our evaluation framework. [1]

[1]Code available at https://github.com/tdw75/beyond-marginal-distributions

Findings of the Association for Computational Linguistics: ACL 2026, pages 4806–4826

Using the World Values Survey (WVS) as our ground truth, we apply this correlation-based framework to compare two methods for model steering (i.e., conditioning models to become more representative of various demographic subgroups): (1) persona prompting with general-purpose LLMs; and (2) demographically fine-tuned models, specifically OpinionGPT (Haller et al., 2024). Persona prompting uses prompt engineering to alter the model's input context and invoke its internal "concept" of a demographic from its pretrained distribution. In contrast, fine-tuning embeds a group's perspectives directly into the model's parameters by training on natural language corpora authored by members of that group.

Our primary contribution is the introduction of a new framework for survey-based evaluations of LLMs that uses both marginal distributions and correlation structures to reveal hidden failures in model representativeness and is easy to adapt to other use cases as required. As a secondary contribution, we demonstrate the diagnostic value of our framework through an illustrative case study comparing demographic fine-tuned and persona-prompted LLMs for model steering.

Using this framework, we investigate the following research questions:

RQ1 To what extent can common steering methods shift marginal response distributions toward those observed in human survey data?

RQ2 Do improvements in marginal alignment imply improved structural alignment in terms of inter-question correlation patterns?

RQ3 How does the representativeness vary across demographic subgroups and value domains when evaluated under this framework?

We find that while fine-tuning and persona prompting can move models towards "human-like" marginals, substantial divergence in the correlation structure remains under both approaches.

2 Related Work

2.1 Representativeness in LLMs

Empirical studies have found LLMs to be more representative of some demographic groups than others. Geographically, they tend to align more closely with Western values (Durmus et al., 2024; Tao et al., 2024). Santurkar et al. (2023) also find representativeness disparities by education, religion, and socioeconomic status, where the opinions of OpenAI's instruction-tuned models align better with those of the annotating group's demographics.

Politically, LLMs have been shown to represent left-leaning opinions (Perez et al., 2023; Hartmann et al., 2023; Martin, 2023; Feng et al., 2023; Fujimoto and Takemoto, 2023; Weeber et al., 2026). Ceron et al. (2024) found that while most models do display *political bias* (the favouring of specific policy positions), there is insufficient evidence of a *political worldview* (a consistent ideological orientation across domains).

Even studies claiming representative results in LLMs should be treated with caution, with many conclusions based on demographically narrow evaluations (Sen et al., 2025). In addition to the general misalignment, the distribution of true human responses shows greater diversity than that of model responses (Durmus et al., 2024; Santurkar et al., 2023). This is particularly a problem in instruction-tuned models, in line with theoretical limitations of RLHF (Kirk et al., 2024; Xiao et al., 2024).

2.2 Measuring Representativeness

Most studies on human-LLM alignment generate responses independently at the survey item level and assess representativeness by comparing the marginal response distributions through statistical metrics (response means or variances) and distance/divergence metrics of answer distributions (Jenson-Shannon, Wasserstein, Kullblack-Leibler, etc.) (Bisbee et al., 2024; Argyle et al., 2023; Santurkar et al., 2023; Durmus et al., 2024; Tao et al., 2024; Sen et al., 2025; Gupta et al., 2024). While these metrics are essential, they treat each question individually.

In contrast, work in the social sciences considers underlying multivariate value patterns to be central. Examples are classic cultural value frameworks such as the Inglehart-Weizel Cultural Map (Inglehart, 2000; Inglehart and Welzel, 2023), Hofstede's Cultural Dimensions (Hofstede, 1980), and Schwartz's Theory of Basic Human Values (Schwartz, 1992). True representativeness therefore also requires the preservation of the underlying structures that constitute cultural dimensions. Recent work by Münker (2025) proposes a similar framework to ours, also based on correlation structure and termed *fingerprinting*, for analysing LLM-generated survey responses demonstrated on a single psychometric instrument. In contrast, our

work sees correlation structures as a necessary criterion for population-level representativeness and value alignment with human survey data.

2.3 Steering and Aligning LLMs

Model steering aims to counteract the collapse toward an "average human preference" observed in many modern LLMs (Sorensen et al., 2024). Instead, they seek to explicitly push model outputs to reflect the viewpoints of a specific demographic, a process aligned with the goal of distributional pluralistic alignment. In the context of this paper, we define ***model steering*** as any method that attempts to encourage a language model's outputs to more closely reflect a specific target distribution. This can be achieved in different ways. We discuss two:

Persona prompting is widely used to steer a model towards the perspective of a demographic subgroup using prompts that define attributes or an identity, i.e., a persona. As a form of prompt engineering as it does not require retraining and can therefore be applied to almost any available LLM without having access to model weights or output probabilities. While there has been evidence of limited steerability toward the intended target groups (Santurkar et al., 2023; Argyle et al., 2023; Bisbee et al., 2024; Durmus et al., 2024), in many cases the resulting distributions could still not be considered to be representative. Outputs fail to capture the intra-group heterogeneity found in true responses (Santurkar et al., 2023; Bisbee et al., 2024) and often rely on reductive, stereotypical or even harmful associations (Durmus et al., 2024; Gupta et al., 2024). Therefore, it is unclear whether persona prompting as a steering technique can truly capture the diversity of human subgroups rather than merely activating surface-level semantic features.

Fine-tuning is an alternative to prompt-based model steering by adjusting the model weights or training an adapter (Hu et al., 2021) on data that is representative of the group the model should be aligned to. *OpinionGPT* (Haller et al., 2024) uses Reddit data to create adapters for different genders, geographic regions, age groups, and political views. To align models with more left- or right-leaning political views, Feng et al. (2023) use data from Reddit and news outlets while Weeber et al. (2026) use political manifestos. *CultureLLM* (Li et al., 2025) and *SubPOP* (Suh et al., 2025) leverage fine-tuning on public opinion surveys for task-specific improvements and improved cultural alignment. Although less flexible than prompting, results from fine-tuning suggest that internalising subgroup perspectives through parameter adaptation may yield more stable, generalisable, and representative models. While fine-tuning requires computational effort and access to model weights, existing models can be reused, as we do in this work.

3 Evaluation Framework

To assess how representative generated responses are of true human responses, our approach considers both the marginal response distributions as well as the correlation structures. Each comparison addresses a different question:

1. ***Marginal distributions:*** How well do simulated responses approximate the ground truth response distribution? How diverse are the outputs compared to true survey responses?
2. ***Correlation structures:*** Do the simulated responses reproduce the interdependencies between questions and topics that underpin the WVS and related social scientific research?

To outline the framework we use a motivating example similar to the experimental setup in similar value-alignment studies (Santurkar et al., 2023; Durmus et al., 2024). Here, we have:

- a set of *survey questions* Q (e.g., the WVS);
- a population with various *demographic subgroups* S (e.g., Germans, women, liberals, etc.) that has responded to each $q \in Q$; and
- a set of *models* M, where each $m \in M$ corresponds to a subgroup $s \in S$ and is used simulate responses to each $q \in Q$.

3.1 Constructing and Aggregating Response Distributions

We first construct the ***ground truth response distribution*** of each demographic subgroup $s \in S$ (for each question $q \in Q$) weighing responses using available survey weights. For each model $m \in M$ corresponding to the subgroup s, we then construct the matching ***simulated response distribution*** using the generated outputs. This yields two distributions (one observed, P_s, and one simulated, P_m) of responses to each question from each subgroup. Figure 2 represents the construction of these marginal response distributions for human and simulated respondents.

To evaluate different groupings, responses can additionally be aggregated along one of the two axes before constructing the distributions: (1) ***demographic dimensions***, e.g., to investigate larger

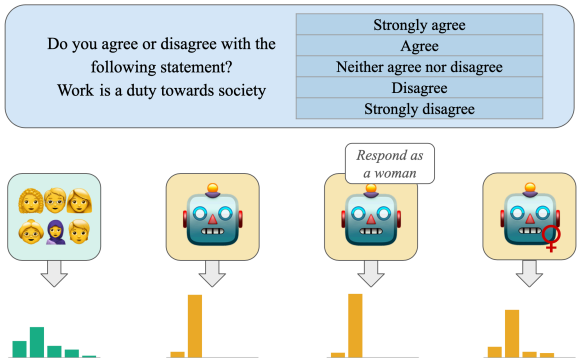

Figure 2: Process for constructing the marginal distributions of human opinions (green) and different steered models (yellow).

subsets of the respondents; and (2) ***question topics***, e.g., to analyse broader value domains.

3.2 Evaluating Marginal Response Distributions

A major quantity of interest is the similarity between the empirical (marginal) response distribution, P_s, and the simulated ones, P_m, of the corresponding demographic subgroup (Durmus et al., 2024; Santurkar et al., 2023). We measure this with the marginal ***dissimilarity*** (on Q) as:

$$\mathcal{D}(P_m, P_s) = \frac{1}{|Q|} \sum_{q \in Q} d(P_m(\cdot \mid q), P_s(\cdot \mid q))$$

where $d(\cdot, \cdot) \in [0, 1]$ is a distance on probability distributions. $\mathcal{D}(P_m, P_s)$ is then the mean of per-question distances, yielding a unit-free dissimilarity score between 0 (perfectly representative) and 1 (maximal divergence).

As a noted weakness of LLM-generated responses (Santurkar et al., 2023; Bisbee et al., 2024; Durmus et al., 2024), we isolate response diversity by comparing the means of the normalised per-question variances for the true and generated responses. The ***mean response variance*** is given by:

$$\mathcal{V}_s(P) = \frac{1}{|Q|} \sum_{q \in Q} \frac{\mathrm{Var}_{r \sim P(\cdot|q)}(r)}{\mathrm{diam}(R_q)^2}$$

where $P(r \mid q)$ denotes the probability assigned to response $r \in R_q$ for question q under distribution $P \in \{P_s, P_m\}$. The variance $\mathrm{Var}_{r \sim P(\cdot|q)}(r)$ is computed over the response values r with respect to this distribution and normalised by the diameter, $\mathrm{diam}(R_q)$, of the response set for question q.

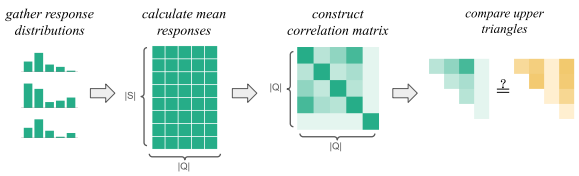

Figure 3: Process for constructing a correlation matrix from human opinions (green) and comparing it with a simulated correlation matrix (yellow).

3.3 Evaluating Correlation Structures

Beyond marginal distributions, representativeness also requires that the simulated data preserve the ***correlation structure*** between questions present in human responses. E.g., if respondents that score highly on *Q1: importance of family* and also score highly on *Q45: respect for authority*, then those items exhibit a positive correlation, potentially reflecting an underlying traditionalist worldview. As illustrated in Figure 3, we compare the correlation structures of empirical and simulated responses as follows:

1. We gather the ***response distributions***, P_s or P_m
2. We compute the [0-1]-normalised ***question mean response*** to item $q \in Q$ from subgroup $s \in S$ under each distribution, yielding two mean matrices $\mathrm{A}^{\mathrm{true}}$ and $\mathrm{A}^{\mathrm{sim}}$, both in $\mathbb{R}^{|S| \times |Q|}$
3. We construct a ***question-question correlation matrix*** $\mathrm{C} \in \mathbb{R}^{|Q| \times |Q|}$ as the pairwise correlation coefficients between columns[2] of A (for calculation details, see Appendix C.4)
4. We compare the empirical correlation matrix $\mathrm{C}^{\mathrm{true}}$ to the simulated one $\mathrm{C}^{\mathrm{sim}}$ in two complementary ways: (1) the ***Pearson correlation***:

$$\rho_{\mathrm{true,sim}} = \mathrm{corr}\left(\mathrm{u}^{\mathrm{true}}, \mathrm{u}^{\mathrm{sim}}\right)$$

as well as (2) the ***Root Mean Square Error (RMSE)*** (for the n unique elements):

$$\mathrm{RMSE}_{\mathrm{true,sim}} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} \left(u_i^{\mathrm{true}} - u_i^{\mathrm{sim}}\right)^2}$$

where u denotes the vector formed by stacking the upper-triangular, off-diagonal entries of a correlation matrix C.

This choice of two metrics[3] (correlation and

[2]To construct a correlation matrix at a different level of aggregation (e.g., a *topic-topic correlation matrix*), the same procedure applies with the extra step of first aggregating the response distributions accordingly.

[3]These metrics are used as a descriptive and comparative measure of structural alignment between vectorised correla-

RMSE) allows us to separate whether a model reproduces the relative *structure* of correlations (i.e., which question pairs tend to move together) from whether it also matches their *magnitude*.

4 Experimental Setup

Using the framework defined in section 3, we now compare persona prompting and fine-tuning as steering methods to make model outputs more representative of demographic subpopulations.

4.1 Models

We consider three model configurations: (1) an unsteered LLM; (2) a persona prompted LLM; and (3) a demographic fine-tuned LLM: OpinionGPT.

(1) Unsteered Baseline `phi-3` (Abdin et al., 2024). In this configuration, we do not try to steer the model to represent a specific demographic. This baseline allows us to assess the improvement of the following two steering methods.

(2) Persona Prompting `phi-3` + persona prompt. A simulation run with the addition of one of ten different demographic-specific profile instructions. This is the steered baseline and main point of comparison for demographic fine-tuning.

(3) Demographic Fine-tuning We use an existing demographically fine-tuned model: OpinionGPT (Haller et al., 2024), which consists of eleven fine-tuned LoRA adapters across four demographic variables (see Table 4 in the Appendix for an overview of adapters). OpinionGPT was created to make socio-demographic biases in LLMs explicit by training each adapter on data generated only by the target demographic subgroup. For this purpose, the authors leveraged data from the *r/AskAXXX* subreddits, e.g., *r/AskAGerman*. We do not include any additional steering in the prompt.

For each steered configuration, we generate responses for ten demographic subgroups: Two ***gender*** attributes (*male*, *female*), two ***age***[4] attributes (*people over 30*, *old people*), four ***geographic*** attributes (*German*, *US American*, *Latin American*, *Middle Eastern*), and two ***political*** attributes (*liberal*, *conservative*). Together with the unsteered baseline, we have 21 different models.

tion matrices rather than for statistical inference, thus, we do not rely on distributional assumptions.

[4]We simulate responses for each demographic subgroup from OpinionGPT except for teenagers as this group is not represented in our evaluation data.

4.2 Evaluation Data

The World Values Survey (WVS) (Inglehart, 2000) is an academic study of social, political, economic, religious, and cultural values and is conducted periodically with a new survey wave released every 5-10 years. The set of categories remains broadly consistent across waves to facilitate longitudinal and cross-national comparisons. Its multiple choice format allows for easy quantitative comparison, which combined with the complex and abstract nature of the survey topics makes it a popular resource for assessing LLM value alignment (Durmus et al., 2024; Li et al., 2025). The close correspondence of the OpinionGPT modules to the data's demographic composition allows a like-for-like comparison for the available subgroups (see Table 1).

OpinionGPT	WVS
Liberal	respondent answered 1, 2, or 3 on a 1-10 scale from political left to right
Conservative	respondent answered 8, 9 or 10 on a 1-10 scale from political left to right
German	respondent from Germany
US American	respondent from the US
Latin America	respondent from any available Latin American country
Middle East	respondent from any available Middle Eastern country
Men	respondent is male
Women	respondent is female
People Over 30	respondents born after 1980 but over 30 years of age
Old People	respondent born in or before 1980 (definition from subreddit)

Table 1: All OpinionGPT demographic subgroups from the political, geographic, gender, and age dimensions and their respective match in the WVS data.

We take survey data from WVS wave 7 (Haerpfer et al., 2020), which organises questions into various topics (see subsection B.1 for a summary). We select a subset of 193 questions that support easy comparison across the ten OpinionGPT subgroups and the use of a standard prompting structure like in similar studies (Santurkar et al., 2023) and QA tasks more generally (Liang et al., 2023). We use the human survey response distributions as a ground-truth to assess LLM representativeness. Figure 2 also shows an example question.

4.3 Prompting and Simulation

For each of the 21 models, we simulate 193 questions with 500 samples per question at a temperature of 0.9 (see Appendix E.1 for temperature analysis). After simulation we identify any responses

that cannot be matched to an admissible response (e.g., due ambiguity or deviation from task) and mark this as invalid, analogous to a non-response in a human survey (see Appendix B for details).

We keep the system prompt simple with a brief description of the task and a single example to define the desired output format (see Appendix B.3). Our persona prompting approach is similar to previous work (Bisbee et al., 2024; Durmus et al., 2024; Santurkar et al., 2023) where a short text is added to the prompts instructing the model to respond from the perspective of the relevant demographic subgroup. The format is shown in Appendix B.3.

4.4 Response Distributions and Aggregation

For each WVS question q, we have 10 subgroup-specific empirical distributions. We construct 21 model response distributions as outlined in Section 4.1 (10 subgroups $\times$ 2 steering strategies plus unsteered baseline). We then aggregate once along each axis (demographics and questions). ***Demographic dimension aggregation*** aggregates the data over all available profiles within a given dimension (e.g. *men* and *women* in the gender dimension) resulting in, for each question q, a distribution for gender, political leaning, geographic/cultural origin, and age (see: Table 5); each can be seen as an approximation with a single response distribution of the entire WVS survey population using available demographic subgroups.

For ***question topic aggregation***, we aggregate all questions from the same thematic category of the WVS, resulting in 12 separate aggregated distributions for each subgroup s or model m. This enables us to evaluate whether models capture the *relationships between broad value domains* rather than just item-level associations.

5 Analysis 1: Marginal Distributions

This first analysis takes the traditional evaluation approach, investigating whether each model configuration can steer model outputs to better match the empirical marginal response distributions.

5.1 Setup

Using the response distributions from the subsection D.3, we follow the procedure outlined in subsection 3.2. Here, we use the diameter-normalised *Wasserstein-1* as our distance metric, d, for questions with an ordered response scale (e.g., Likert-type or numerical ratings), and the *total variation* distance for questions with a nominal response set. See Appendix C.3 for details.

5.2 Results

The left-hand panel in Figure 4 plots dissimilarity scores. It shows that model steering (through either OpinionGPT or persona prompting) produces more representative marginal response distributions than the unsteered baseline model. OpinionGPT reduces mean dissimilarity over the unsteered baseline for every subgroup, as does persona prompting for all but the *liberal* demographic. The improvement from OpinionGPT is greater than that from persona prompting for all but two demographic subgroups.

Similar patterns can be seen when aggregating along the demographic dimensions (Figure 4, bottom left). OpinionGPT and persona prompting improve noticeably over the baseline for each dimension, with large improvements for OpinionGPT compared to persona prompting across all dimensions. However, markedly different patterns emerge with question topic aggregation (see Appendix Figure 7). Both OpinionGPT and persona prompting reduce mean dissimilarity relative to the unsteered baseline, but the extent of improvement varies substantially across value domains. The right-hand panel of Figure 4 shows the mean response variance by demographic subgroup for each model configurations and the true responses. In each subgroup, OpinionGPT induces more response diversity to the responses than the unsteered baseline, whereas persona prompting decreases it. One reason for the suppressed variance is the collapse to a degenerate response distribution, which happens more frequently for persona prompting (see Appendix D.2). An exception is the *political leaning* subgroups and dimension where all model configurations suppress response diversity to below empirical levels.

6 Analysis 2: Correlation Structures

Our first analysis found promising results in the models' ability to approximate marginal response distributions of the WVS. We now move beyond marginals to assess whether they can also produce the latent structures underpinning cultural theories (Hofstede, 1980; Schwartz, 1992).

6.1 Setup

We analyse correlation matrices at two levels: (i) the ***question–question*** level, which considers

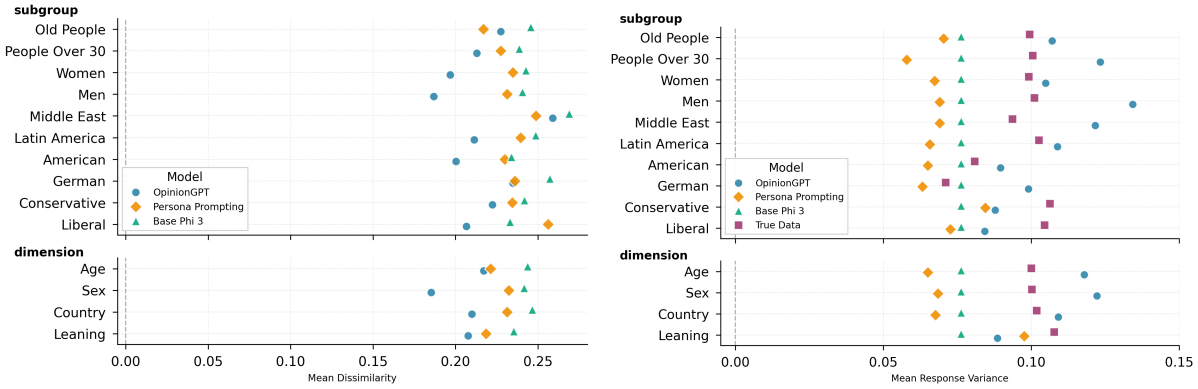

Figure 4: Evaluation metrics by demographic subgroup and dimension. Left: mean dissimilarity (lower = better). Right: mean response variance (closer to true data = better).

correlations between individual survey questions[5]; and (ii) the ***topic–topic*** level, where questions are collapsed into the 12 thematic value domains in our WVS subset. The first tests whether models capture fine-grained associations between specific attitudes, while the second evaluates whether they reproduce broad inter-domain dependencies.

We construct and compare *question-question correlation matrices* as in subsection 3.3. To construct the *topic-topic correlation matrix*, the same procedure applies with the extra step of first aggregating responses to the thematic domains. To account for sampling variability in the simulated responses, we report bootstrapped 95% confidence intervals for each metric (RMSE and Pearson ρ) computed from 500 resamples (with replacement) of the 500 generated responses for each survey item from each model. For full calculation details of the matrices and bootstrap, see Appendix C. As we use question means in this experiment to construct the correlation matrix, we leave out the 18 categorical questions, which lack the notion of a mean. This omits $<$10% of the 193 questions, more details in Appendix D.3.

We further contextualise the proposed correlation-structure metrics by estimating lower and upper bounds on model performance via a ***permutation-based null*** baseline and a ***split-half resampling*** of the WVS, respectively.

The permutation-based null provides a baseline[6] corresponding to the absence of cross-question correlation between subgroups. We construct this by independently permuting the mean WVS responses (each column of $\mathrm{A}^{\mathrm{WVS}}$) before recalculating the correlation matrix and comparing this to the true $\mathrm{C}^{\mathrm{WVS}}$ using the same evaluation metrics. This preserves the marginal distribution of response means for each survey item while destroying correlations between items.

The *split-half resampling* complements this by providing an empirical ceiling for the metrics through an estimate for the most optimistic case, i.e., when the correlation matrices are constructed from subsets of the exact same distribution of responses. To perform this, we randomly partition WVS responses into two halves, computing correlation matrices for each subset and again comparing them using the same evaluation metrics. Each is performed $1,000$ times with the percentile values reported corresponding to a 95% interval. Details can be found in Appendix C.6.

6.2 Results

Relative to a structureless permutation-based null baseline and a near-noiseless split-half resampling, the steering methods do recover a degree of the inter-subgroup correlation structure at the question-question level (see Table 2). However, both OpinionGPT and persona prompting still remain well below the empirical ceiling, indicating poor recovery of fine-grained dependence patterns. While neither model accurately reproduces the *relative pattern* of correlations, persona prompting does so noticeably better than OpinionGPT (ρ: 0.158 vs 0.09). The respective bootstrapped intervals (see Table 2) are tight around the point estimates, indicating that the effect of sampling variability of these metrics is fairly low. Despite displaying stronger directional similarity than OpinionGPT,

[5]This calculation applies only to ordinal questions $Q^{ord} \subset Q$, omitting the $<$10% nominal-answer questions.

[6]Our baseline model is demographic-agnostic, this lack of this axis means there are is no subgroup correlation structure. The permutation-based null serves as a baseline instead.

Model	Pearson ρ	RMSE
OpinionGPT	0.090[0.08,0.10]	0.638[0.63,0.64]
Persona	0.158[0.15,0.17]	0.679[0.67,0.68]
Perm. Null	−0.004	0.849
Split Half	0.999	0.006

Table 2: Question-question correlation metrics with 95% bootstrapped confidence intervals in brackets

Model	Pearson ρ	RMSE
OpinionGPT	−0.018[-0.02,0.05]	0.718[0.71,0.73]
Persona	0.240[0.21,0.28]	0.676[0.67,0.69]
Perm. Null	0.001	0.914
Split Half	0.997	0.011

Table 3: Topic-topic correlation metrics with 95% bootstrapped confidence intervals in brackets

from the RMSE we can see that persona prompting produces correlation patterns that are further away in *magnitude* (RMSE: 0.679 vs 0.638). Again, the tight confidence intervals seen in Table 2 are tight around the point estimates.

When aggregating items into question topics, the preservation of correlation structures from OpinionGPT degrades notably (see Table 3). With ρ: -0.018 the relative pattern of correlations is now completely lost; the RMSE is also higher in comparison to the question-question case. Conversely, persona prompting improves on both metrics after aggregation (ρ: 0.24, RMSE: 0.676). Once again, the confidence intervals are narrow around the point estimates. Although the results in section 5 show that OpinionGPT produces marginal distributions that are much closer to the WVS ground truth, the results of this analysis show that persona prompting performs marginally better at reproducing the empirical correlation structure across subgroups. Nevertheless, neither technique faithfully reproduces these correlation patterns.

7 Discussion

Representativeness of Marginal Distributions. We find mixed results for our RQ1: OpinionGPT improved marginal distribution similarity over the unsteered baseline across all demographics (see Figure 4). Improvements were also partially evident for persona prompting, in line with previous findings (Bisbee et al., 2024; Santurkar et al., 2023). (Leidinger et al., 2023) or algorithmic search (Zheng et al., 2024). Yet, while the relative improvements were far more consistent and substantial for OpinionGPT, distributional dissimilarity is not eliminated and remains uneven and context-dependent; subgroups such as *Middle East* and *old people* are less well represented, potentially due to poor representation in the Reddit fine-tuning data.

Response diversity has been noted as a particular weakness with LLM responses (Santurkar et al., 2023; Bisbee et al., 2024; Durmus et al., 2024). As seen in Figure 4, OpinionGPT induces greater response diversity compared to the unsteered baseline, which in line with previous findings is already less variant than true responses. However, the variance of OpinionGPT responses often exceeds empirical levels, producing distributions that overstate the degree of disagreement within a demographic group. Persona prompting displays the opposite problem and greatly suppresses response diversity. The mechanisms of the respective steering methods differ substantially. Fine-tuning reshapes model parameters using data directly from the target demographic and in doing so appears to induce more stable and heterogeneous marginal response distributions. In contrast, persona prompting functions as a strong conditioning signal that narrows the effective response space. This signal often collapses the distribution toward a single stereotypical response (see Appendix Table 9 for more detail), completely suppressing the heterogeneity observed in the WVS; this is a commonly highlighted issue with persona prompted models (Durmus et al., 2024; Gupta et al., 2024).

Correlation Structures and Aggregation Effects. Previous findings suggest that steered or aligned LLMs are better at capturing surface-level alignment patterns than deep latent structure (Santurkar et al., 2023; Durmus et al., 2024; Gupta et al., 2024; Kabir et al., 2025). Using our framework to answer RQ2, we add further evidence to this, emphasising that improved marginal distribution similarity through model steering does not imply the preservation of latent structures. While persona prompting performs slightly better than OpinionGPT, neither model is able to adequately able to capture question-question correlation structure of WVS data (Table 2) or the topic–topic correlation (Table 3). This highlights a key weakness: Simulated responses can approximate marginal response distributions to an extent but have more difficulty reproducing the higher-order relationships between value domains that constitute a coherent political worldview (Ceron et al., 2024) and, which give sur-

veys such as the WVS their interpretive coherence. Again, the differing steering mechanisms may offer one possible explanation. OpinionGPT uses separate adapters for each demographic group, which may allow representations to drift across groups. In contrast, persona prompting relies on a single conditioned model, meaning that responses across groups are generated from a shared underlying representation. The greater heterogeneity previously observed for OpinionGPT may therefore help reproduce diverse marginal response distributions, but could be disadvantageous when modelling the relational structure across demographic subgroups.

Finally, RQ3 examines whether the representativeness observed in simulated responses persists when results are aggregated. The results suggest that aggregation by demographic dimension does not substantially diminish representativeness, preserving marginal similarity and variance patterns. In contrast, aggregation by question topic produced less stable results, with marginal similarity varying substantially across domains, in addition to the above mentioned issues with correlation structure preservation. This suggests that representativeness depends not only on demographic context, but also on the question topic. Despite some limitations, OpinionGPT better approximates marginal response distributions than the unsteered baseline across value domains and, with the exception of *Ethical Values*, outperforms or roughly equals persona prompting. However, after the same aggregation to topic level the correlation structure of OpinionGPT responses loses all relative similarity to the true WVS, whereas this slightly improves for persona prompting, despite remaining poor overall.

Broader Implications for Representativeness. The findings underscore that representativeness constitutes a distinct axis of model alignment, separate from established dimensions such as safety, helpfulness, or factuality (Gabriel, 2020). These can often be assessed at the level of individual responses, but representativeness is inherently defined at the distributional level, requiring the preservation not only of central tendencies but also of variance, demographic fidelity, and correlation structures (Santurkar et al., 2023; Argyle et al., 2023; Dominguez-Olmedo et al., 2024). Our results demonstrate that our tested approaches lack this representativeness. Therefore, long-standing cultural values theories such as those of Hofstede and Schwartz, which show that meaningful cultural differences emerge from latent structures derived from multivariate correlation in survey data rather than from any single attitude in isolation, do not hold for response from our tested LLMs. The results from our empirical case study highlight the importance of evaluating both marginal response distributions and inter-item correlations when assessing demographic alignment. Our framework enables this joint evaluation and reveals differences between steering approaches that would be overlooked by analyses focusing on marginals alone.

8 Conclusion

In this paper, we have argued that representativeness is a distinct and necessary dimension of alignment, one that requires the preservation of not only diversity and subgroup fidelity but also the structural interdependencies that characterise human populations and underpin major cultural values frameworks in the social sciences. We therefore propose an evaluation framework that explicitly considers both marginal response distributions and inter-question correlation structures. Applying this to the WVS, we tested two model steering approaches for demographic alignment: (1) demographic fine-tuning, with OpinionGPT; and (2) persona prompting, with Phi-3-Mini-Instruct-4k. Our results show that while OpinionGPT shows promising results in marginal similarity and variance structure when compared to persona prompting, it performs worse at reproducing the correlation structures, although neither steering approach was able to faithfully preserve these higher order interdependencies. We thus emphasise the fragility of demographic alignment and the danger of making representativeness claims based on marginal distribution properties alone.

An important direction for future work is the incorporation of population-level dependency structures into alignment mechanisms to enable both steerability across groups as well as fidelity to population-wide distributions, especially regarding the interconnectedness of how values and opinions are structured.

Limitations

Model limitations. OpinionGPT's base model, *Phi 3 Mini 4k*, is a compact LLM with far fewer parameters and a shorter context window than frontier-scale models. Through targeted data curation, the authors nonetheless achieve strong per-

formance, often comparable to larger models (Gunasekar et al., 2023; Abdin et al., 2024), but also note size-related limits that may reduce the ability to capture complex demographic variation. Smaller models are more prone to deviating from instructions and prescribed response formats, adding noise and errors (Murthy et al., 2024). OpinionGPT was also fine-tuned on Reddit data, whose users skew young (Pew Research Center, 2024) and whose lingua franca is English rather than region-specific languages, limiting representativeness via the fine-tuning corpora. Finally, focusing on a single model is itself a limitation. Our intent is to present a nuanced evaluation framework, which future work could extend to other alignment methods and base models.

Design choices. We independently sample 500 responses per item (for each model configuration) and compare distributions. While common in simulated survey alignment work (Santurkar et al., 2023; Durmus et al., 2024), this requires aggregating by demographic subgroup to construct correlation matrices (see subsection C.4), losing some distributional information. A trajectory-based alternative would condition each answer on prior ones to generate full respondent surveys from which correlation matrices could be constructed. This would allow respondent-level analyses and finer-grained item correlations while remaining compatible with our framework. In subsection 3.3, one would skip mean-response computation in step 2 and construct correlations in step 3 directly from trajectories.

WVS and multiple choice surveys. The WVS, rooted in a European values project, embeds Western and Eurocentric normative assumptions (Goodwin et al., 2020), shaping both question design and interpretation and making representativeness relative to a culturally contingent benchmark rather than a neutral truth. Our analysis is also English-only, so we cannot claim generalisability to other languages. For geographic subgroups, prompting in region-native languages (e.g., Arabic for the Middle East) might yield better alignment than English, given likely regional differences in pre-training data sources. Multiple-choice surveys further offer clear advantages such as straightforward encoding, comparable probability distributions, and large-scale robust resources, hence their use in survey research and LLM alignment (Santurkar et al., 2023; Durmus et al., 2024). However, the closed-ended tasks necessarily restrict nuance and may miss misalignment that appears in open-ended settings.

Demographic Subgroups. We evaluate ten demographic subgroups defined by OpinionGPT adapters, which do not exhaust relevant respondent characteristics. Coverage is limited to male/female genders, two US-spectrum political views, and no people below 30, while geographic groups differ in granularity (single-country groups like Germans or US Americans versus broader regions like Middle Easterners or Latin Americans spanning diverse contexts). Other dimensions (e.g., socio-economic status or education) may also matter, and our one-dimensional subgrouping ignores intersectional perspectives.

Ethics statement

All our data is publicly available and licensed for research. We did not fine-tune any models ourselves, but the models we use have been fine-tuned on publicly available data only. The WVS includes real data from survey respondents, but they have been anonymised before the publication of the WVS data. We also caution against conflating high scores on specific representativeness metrics with genuine model alignment.

Acknowledgements

Alan Akbik is supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Emmy Noether grant "Eidetic Representations of Natural Language" (project number 448414230). Further, Alan Akbik is supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy "Science of Intelligence" (EXC 2002/1, project number 390523135).

References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Qin Cai, Vishrav Chaudhary, Dong Chen, Dongdong Chen, and 110 others. 2024. Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone. *arXiv preprint*. Version Number: 4.

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.

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.

Tanise Ceron, Neele Falk, Ana Barić, Dmitry Nikolaev, and Sebastian Padó. 2024. Beyond Prompt Brittleness: Evaluating the Reliability and Consistency of Political Worldviews in LLMs. *Transactions of the Association for Computational Linguistics*, 12:1378–1400.

Ricardo Dominguez-Olmedo, Moritz Hardt, and Celestine Mendler-Dünner. 2024. Questioning the Survey Responses of Large Language Models. *Advances in Neural Information Processing Systems*, pages 45850–45878.

Esin Durmus, Karina Nguyen, Thomas 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. 2024. Towards Measuring the Representation of Subjective Global Opinions in Language Models. In *Proceedings of COLM*.

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. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11737–11762, Toronto, Canada. Association for Computational Linguistics.

Sasuke Fujimoto and Kazuhiro Takemoto. 2023. Revisiting the political biases of ChatGPT. *Frontiers in Artificial Intelligence*, 6:1232003.

Iason Gabriel. 2020. Artificial Intelligence, Values, and Alignment. *Minds and Machines*, 30(3):411–437.

Jamie Lynn Goodwin, Andrew Lloyd Williams, and Patricia Snell Herzog. 2020. Cross-Cultural Values: A Meta-Analysis of Major Quantitative Studies in the Last Decade (2010–2020). *Religions*, 11(8):396. Multidisciplinary Digital Publishing Institute.

Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio César Teodoro Mendes, Allie Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero Kauffmann, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Harkirat Singh Behl, Xin Wang, Sébastien Bubeck, Ronen Eldan, Adam Tauman Kalai, Yin Tat Lee, and Yuanzhi Li. 2023. Textbooks Are All You Need. *arXiv preprint*. Version Number: 2.

Shashank Gupta, Vaishnavi Shrivastava, Ameet Deshpande, Ashwin Kalyan, Peter Clark, Ashish Sabharwal, and Tushar Khot. 2024. Bias Runs Deep: Implicit Reasoning Biases in Persona-Assigned LLMs. In *The Twelfth International Conference on Learning Representations*. ArXiv:2311.04892 [cs].

Christian Haerpfer, Ronald Inglehart, Alejandro Moreno, Christian Welzel, Kseniya Kizilova, Jaime Diez-Medrano, Marta Lagos, Pippa Norris, Eduard Ponarin, and Bi Puranen. 2020. World Values Survey Wave 7 (2017-2020) Cross-National Data-Set.

Patrick Haller, Ansar Aynetdinov, and Alan Akbik. 2024. OpinionGPT: Modelling Explicit Biases in Instruction-Tuned LLMs. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 3: System Demonstrations)*, pages 78–86, Mexico City, Mexico. Association for Computational Linguistics.

Jochen Hartmann, Jasper Schwenzow, and Maximilian Witte. 2023. The political ideology of conversational AI: Converging evidence on ChatGPT's pro-environmental, left-libertarian orientation. *arXiv preprint*. ArXiv:2301.01768 [cs].

Geert Hofstede. 1980. *Culture's Consequences: International Differences in Work-Related Values*. Sage.

Allyson L. Holbrook, Jon A. Krosnick, David Moore, and Roger Tourangeau. 2007. Response Order Effects in Dichotomous Categorical Questions Presented Orally: The Impact of Question and Respondent Attributes. *The Public Opinion Quarterly*, 71(3):325–348. Publisher: [Oxford University Press, American Association for Public Opinion Research].

Geoff Hollis, Chris Westbury, and Lianne Lefsrud. 2017. Extrapolating human judgments from skip-gram vector representations of word meaning. *Quarterly Journal of Experimental Psychology*, 70(8):1603–1619.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The Curious Case of Neural Text Degeneration. In *8th International Conference on Learning Representations (ICLR 2020)*. ArXiv:1904.09751 [cs].

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. In *10th International Conference on Learning Representations*.

Ronald Inglehart. 2000. Globalization and postmodern values. *The Washington Quarterly*, 23(1):215–228.

Ronald Inglehart and Christian Welzel. 2023. The Inglehart-Welzel World Cultural Map. *World Values Survey 7*.

Shariar Kabir, Kevin Esterling, and Yue Dong. 2025. Testing Conviction: An Argumentative Framework for Measuring LLM Political Stability. *arXiv preprint*. ArXiv:2504.17052 [cs].

Robert Kirk, Ishita Mediratta, Christoforos Nalmpantis, Jelena Luketina, Eric Hambro, Edward Grefenstette, and Roberta Raileanu. 2024. Understanding the Effects of RLHF on LLM Generalisation and Diversity. In *12th International Conference on Learning Representations*.

Jon A. Krosnick and Duane F. Alwin. 1987. An Evaluation of a Cognitive Theory of Response-Order Effects in Survey Measurement. *Public Opinion Quarterly*, 51(2):201.

Alina Leidinger, Robert Van Rooij, and Ekaterina Shutova. 2023. The language of prompting: What linguistic properties make a prompt successful? In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9210–9232, Singapore. Association for Computational Linguistics.

Cheng Li, Mengzhuo Chen, Jindong Wang, Sunayana Sitaram, and Xing Xie. 2025. CultureLLM: incorporating cultural differences into large language models. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, pages 84799–84838, Red Hook, NY, USA. Curran Associates Inc.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin New-Man, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, and 31 others. 2023. Holistic Evaluation of Language Models. *Transactions on Machine Learning Research*, 2023-August.

Christina Lu and Max Van Kleek. 2024. Model Plurality: A Taxonomy for Pluralistic AI. In *Pluralistic Alignment Workshop at NeurIPS 2024*.

John Levi Martin. 2023. The Ethico-Political Universe of ChatGPT. *Journal of Social Computing*, 4(1):1–11.

Rudra Murthy, Prince Kumar, Praveen Venkateswaran, and Danish Contractor. 2024. Evaluating the Instruction-following Abilities of Language Models using Knowledge Tasks. *arXiv preprint*. ArXiv:2410.12972 [cs] version: 1.

Simon Münker. 2025. Fingerprinting LLMs through Survey Item Factor Correlation: A Case Study on Humor Style Questionnaire. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 245–258, Suzhou, China. Association for Computational Linguistics.

Ethan Perez, Sam Ringer, Kamile Lukosiute, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, Andy Jones, Anna Chen, Benjamin Mann, Brian Israel, Bryan Seethor, Cameron McKinnon, Christopher Olah, Da Yan, Daniela Amodei, and 44 others. 2023. Discovering language model behaviors with model-written evaluations. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13387–13434, Toronto, Canada. Association for Computational Linguistics.

Pew Research Center. 2024. Social Media Fact Sheet.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose Opinions Do Language Models Reflect? In *Proceedings of the 40th International Conference on Machine Learning*, pages 29971–30004. PMLR. ISSN: 2640-3498.

Shalom H. Schwartz. 1992. Universals in the Content and Structure of Values: Theoretical Advances and Empirical Tests in 20 Countries. In *Advances in Experimental Social Psychology*, volume 25, pages 1–65. Elsevier.

Indira Sen, Marlene Lutz, Elisa Rogers, David Garcia, and Markus Strohmaier. 2025. Missing the Margins: A Systematic Literature Review on the Demographic Representativeness of LLMs. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 24263–24289, Vienna, Austria. Association for Computational Linguistics.

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. 2024. Position: a roadmap to pluralistic alignment. In *Proceedings of the 41st International Conference on Machine Learning*, pages 46280–46302, Vienna, Austria. JMLR.org.

Joseph Suh, Erfan Jahanparast, Suhong Moon, Minwoo Kang, and Serina Chang. 2025. Language Model Fine-Tuning on Scaled Survey Data for Predicting Distributions of Public Opinions. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 21147–21170, Vienna, Austria. Association for Computational Linguistics.

Yan Tao, Olga Viberg, Ryan S Baker, and René F Kizilcec. 2024. Cultural Bias and Cultural Alignment of Large Language Models. *PNAS Nexus*, 3(9):pgae346.

Cédric Villani. 2009. The Wasserstein distances. In Cédric Villani, editor, *Optimal Transport: Old and New*, pages 93–111. Springer, Berlin, Heidelberg.

Franziska Weeber, Tanise Ceron, and Sebastian Padó. 2026. Do political opinions transfer between western languages? an analysis of unaligned and aligned multilingual LLMs. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 319–340, Rabat, Morocco. Association for Computational Linguistics.

Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia

Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, Courtney Biles, Sasha Brown, Zac Kenton, Will Hawkins, Tom Stepleton, Abeba Birhane, Lisa Anne Hendricks, Laura Rimell, William Isaac, and 4 others. 2022. Taxonomy of Risks posed by Language Models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, pages 214–229, New York, NY, USA. Association for Computing Machinery.

Jiancong Xiao, Ziniu Li, Xingyu Xie, Emily Getzen, Cong Fang, Qi Long, and Weijie J. Su. 2024. On the algorithmic bias of aligning large language models with rlhf: Preference collapse and matching regularization. *Journal of the American Statistical Association*, 0(ja):1–21.

Mingqian Zheng, Jiaxin Pei, Lajanugen Logeswaran, Moontae Lee, and David Jurgens. 2024. When "A Helpful Assistant" Is Not Really Helpful: Personas in System Prompts Do Not Improve Performances of Large Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15126–15154, Miami, Florida, USA. Association for Computational Linguistics.

A OpinionGPT

OpinionGPT is a demographically fine-tuned model (Haller et al., 2024) that aims to make biases explicit and transparent rather than attempting to eliminate or obscure them. Low rank adaptation (LoRA) to fine-tune a set of modules, each corresponding to a particular demographic profile enabling users to generate responses from the perspective of each of these groups. In this way, OpinionGPT facilitates controlled exploration of how demographic factors may influence language model outputs, thereby attempting to address the issue of hidden or unacknowledged biases. To construct the demographic LoRA modules, the authors identified a set of subreddits from the *AskX* schema, each associated with a particular demographic subgroup. The full list of demographic modules, their corresponding subreddits, and the sample sizes is provided in Table 4.

As briefly described in Table 5, we have eleven OpinionGPT adapters, of which we consider ten for this analysis. We exclude the teenagers subgroup since there is no good match in the WVS data: While teenagers range from ages 13 to 19, the WVS has respondents age 18 and older only. For all other adapters, we identify the relevant sociodemographic questions and filter all responses based on these questions. Table 5 shows all OpinionGPT subgroups, their corresponding reddit channels and additional channel definitions, and the WVS questions and values we used for filtering.

Demographic	*Source Subreddit*	*Sample Size*
Geographical		
German	AskAGerman	11k
American	AskAnAmerican	20k
Latin American	AskLatinAmerica	20k
Middle East	AskMiddleEast	20k
Political		
Liberal	AskALiberal	20k
Conservative	AskConservatives	18k
Gender		
Female	AskWomen	20k
Male	AskMen	20k
Age Demographics		
Teenager (girls)	AskTeenGirls	10k
Teenager (boys)	AskTeenBoys	10k
Over 30 (men)	AskMenOver30	10k
Over 30 (women)	AskWomenOver30	10k
Old people	AskOldPeople	15.5k

Table 4: List of all demographic LoRA modules and their corresponding subreddits (Haller et al., 2024)

On the political dimension, reddit channels target Liberals and Conservatives, which might be more targeted to the US context, while the question we use operates on a left/right dimension, which is more universal. We still only use responses that show a clear left or right leaning stance.

For the geographic adapters, we used the definitions of the Middle East and Latin America from Wikipedia[7]. While *America* could refer to multiple countries, the channel logo (Uncle Sam with elements of the US flag) clearly refers to the US only.

We also define the age groups to be distinct, i.e., we use the definition of *old people* from the reddit channel and we define *people over 30* to be at least 30, but younger than the *old people* group.

B Data, Simulation and Prompting

B.1 WVS Data Subset

We select a subset of WVS questions to enable easy comparison and a consistent prompting approach. We consider a question "easily comparable" if it (a) is not country-specific, allowing comparison across all geographic subgroups; and (b) does not depend on other questions, enabling an easy prompting structure (as used in similar works (Santurkar et al., 2023) as well as for QA tasks more generally (Liang et al., 2023)) without

[7] https://en.wikipedia.org/wiki/Middle_East, https://en.wikipedia.org/wiki/Latin_America

Demographic Dimension	OpinionGPT Subgroup	Reddit Channel	WVS Question	WVS Question Values for Subgroup
Political	Liberal	AskALiberal	Q240: In political matters, people talk of "the left" and "the right." How would you place your views on this scale, generally speaking? 1 (left) to 10 (right)	1-3 left
	Conservative	AskConservatives		8-10 right
Geographic	German	AskAGerman	Q266 In which country were you born?	Germany
	America	AskAnAmerican		US
	Latin American	AskLatinAmerica (Latin America and the Caribbean)		Argentina, Bolivia, Brazil, Chile, Colombia, Ecuador, Guatemala, Mexico, Nicaragua, Peru, Puerto Rico, Uruguay, Venezuela
	Middle Eastern	AskMiddleEast (Middle East and North Africa)		Cyprus, Egypt, Iran, Iraq, Jordan, Lebanon, Turkey
Gender	Men	AskMen	Q260: Respondent's sex	1 male
	Women	AskWomen		2 female
Age	Teenagers	AskTeenGirls / AskTeenBoys	Q261: Can you tell me your year of birth? Q262: This means you are XX years old?	N/A, excluded because WVS contains ages 18+ only
	People over 30	AskWomenOver30 / AskMenOver30		born after 1980, age over 30
	Old People	AskOldPeople (born in or before 1980)		born in or before 1980

Table 5: All OpinionGPT demographic subgroups from the political, geographic, gender, and age dimensions, the wording of the filter question, and which values defined the subgroup.

the need to reword the question. A summary of included questions by topic can be seen in Table 6.

Topic	Count In Complete Survey	Count In Our Subset
Social values, norms, stereotypes	45	22
Economic values	6	6
Perceptions of corruption	9	9
Perceptions of migration	10	10
Perceptions of security	21	21
Index of postmaterialism	6	0
Perceptions about science and technology	6	6
Religious values	12	11
Ethical values	23	22
Political interest and political participation	36	19
Political culture and political regimes	25	25
Happiness and wellbeing	11	11
Social capital, trust and organizational membership	49	31

Table 6: Summary of included WVS questions by topic

System Prompt

```
You are taking part in a survey on personal
values and will be asked a range of
multiple choice questions.
Each question will have the following
format:

"
Q33: For each of the following statements
I read out, can you tell me how much you
agree with each. Do you agree strongly,
agree, disagree, or disagree strongly? -
Carrot cake is tastier than chocolate cake
Available responses:
1: Agree strongly
2: Agree
3: Disagree
4: Strongly disagree
"

For each question, just select a single
answer from the set of available responses.
Please copy the answer exactly as written
in the 'Available responses' list above.
Do not change spelling or capitalisation.
For example:
4: Agree strongly
```

Figure 5: System prompt used for all simulations

B.2 Simulation Procedure

We use the following process to extract valid outputs and create response distributions for each question from each simulated subgroup.

- For each model (21 in total), we ***initialise a new instance***
- We then ***present each survey item*** (193 in total) to the model as a stand-alone prompt
- For each question (with each model) we ***sample 500 responses*** (temperature 0.9)
- We post-process raw generations to adhere to ***desired response format***. This involves stripping extraneous text, mapping the output to the corresponding numeric response option, and discarding invalid generations that could not be reliably matched
- Finally, we aggregate cleaned responses to ***form response distributions*** for each question and subgroup

B.3 Prompts

The system prompt used for all simulations is given in Figure 5. For persona prompting, each subgroup was instantiated using a single natural-language persona description from the template in Figure 6 using the descriptions in Table 7. Each persona prompt was then appended to the system prompt for the corresponding model. While persona prompting is widely used to condition LLM responses on demographic attributes, the outputs of large language models are known to be sensitive to prompt wording. At present, there is no clear consensus on how to systematically determine optimal prompts, either through linguistic analysis of prompt formulations (Leidinger et al., 2023) or through automated search methods (Zheng et al., 2024). Consequently, we employ a fixed prompt template and subgroup descriptions to ensure reproducibility.

Persona Prompt Template

```
Please answer all questions as if you
are {subgroup_description}. Make sure your
responses align as closely as possible
to the responses that someone from this
demographic is likely to give.
```

Figure 6: Template for the persona prompts.

Subgroup	Persona Description
American	a person born and raised in the United States of America
German	a person born and raised in the Federal Republic of Germany
Latin America	a person born and raised in a Latin American country
Middle East	a person born and raised in a Middle Eastern country
Men	a person that identifies as male
Women	a person that identifies as female
People over 30	a person aged between 30 and 45
Old people	a person aged over 45
Conservative	a person with a conservative or politically right-wing view of the world
Liberal	a person with a progressive or politically left-wing view of the world

Table 7: Subgroup descriptions used to instantiate the persona prompt template.

C Evaluation

C.1 Response distributions

We construct response distributions for a given set of questions Q, subgroups S and, models M as follows:

- For each *subgroup* $s \in S$, the survey-weighted ***ground truth response distribution*** for a question $q \in Q$ with possible responses $r \in R_q$ is given by:

$$P_s(r \mid q) = \frac{\sum_{i=1}^{n_{s|q}} w_i \, \mathbf{1}\{ x_{i|q,s}^{\mathrm{WVS}} = r \}}{\sum_{i=1}^{n_{s|q}} w_i}$$

  where:
  - $x_{i|q,s}^{\mathrm{WVS}}$ denotes the response to question q from a WVS respondent $x_i \in X^{\mathrm{WVS}}$ that belongs to subgroup $s \in S$
  - $n_{s|q}$ is the number of observed WVS responses for item q from subgroup s; and
  - w_i denotes the survey weight assigned to respondent $x_i \in X^{\mathrm{WVS}}$.

- For each *model* $m \in M$, the ***simulated response distribution*** for a question $q \in Q$ with possible responses $r \in R_q$ is given by:

$$P_m(r \mid q) = \frac{1}{n_{m|q}} \sum_{i=1}^{n_{m|q}} \mathbf{1}\{x_{i|q}^{m} = r\}$$

  where:
  - $x_{i|q}^{m}$ denotes the simulated response to question q a model m; and
  - $n_{m|q}$ is the number of samples drawn from model m for question q.

C.2 Result Aggregation

To evaluate the robustness of the models, we compare results not only by question for each model configuration, but additionally by aggregating along two different axes corresponding to (1) the *demographics* (and by extension the models); and (2) the *survey questions*.

The two levels of ***demographic aggregation*** align with the OpinionGPT modules. *Demographic subgroup aggregation* aggregates responses over each of the subgroups corresponding to the OpinionGPT modules before calculating an evaluation metric. This isolates each separate simulation run, focusing on the effectiveness of the individual models. *Demographic dimension aggregation* groups the data across all available subgroups in a given dimension (see Appendix Table 4 for list of demographic subgroups and dimensions) before metric calculation; this can be seen as an approximation to the entire WVS survey population.

For ***survey level aggregation*** we collapse survey items into the thirteen thematic categories of the WVS. In addition to the *individual question*, where each question's response set for each question are analysed separately, providing the *most fine-grained view* of model alignment and capturing local covariances between questions, we also aggregate along *question topic*, where item responses are aggregated within the questions categories across all demographic subgroups. Analyses at this level enable us to evaluate whether models capture the *relationships between broad value domains* rather than only item-level associations, and therefore whether representativeness extends to the higher-order structures that underpin social scientific cultural theories.

C.3 Metrics for Marginal Similarity

Wasserstein distance measures the dissimilarity between two probability distributions by measuring the minimal cost of reconfiguring the mass of one distribution such that you recover the other (Villani, 2009). Unlike divergence metrics such as Jensen-Shannon, it takes order into account making it particularly suited to the Likert-type questions common in the WVS. However, it is not suitable for non-ordinal distributions and the unmodified score is not bounded and therefore might be more difficult to interpret.

In the one-dimensional, discrete case the Wasser-

stein distance (for each $q \in Q^{\text{ord}}$) is given by:

$$W(P_m, P_s) = \sum_{r \in R_q} \|F_m(r) - F_s(r)\|$$

where F_m and F_s are the corresponding empirical CDFs to P_m and P_s.For improved interpretability and consistent comparison between questions (and their differing response sets), we use the following normalised version of the Wasserstein distance (similar to that in Santurkar et al. (2023)):

$$d_{\text{W}}(P_m, P_s) := \frac{W(P_m, P_s)}{\text{diam}(R_q)} \in [0, 1]$$

where $\text{diam}(R_q)$ is used as the normalising factor and represents the diameter of the shared support of P_m and P_s for a question $q \in Q^{\text{ord}}$, i.e., the greatest possible distance between any two responses $r, r' \in R_q$

$$\begin{aligned}\text{diam}(R_q) &= \sup_{r,r' \in R_q} |r - r'| = \max_{r,r' \in R_q} |r - r'| \\ &= \max(R_q) - \min(R_q).\end{aligned}$$

Total variation distance measures the largest absolute difference between the probabilities that the two distributions assign to the same event. It is applicable to all discrete distributions, whether ordinal or categorical and normalised and symmetric by default.

For each $q \in Q^{\text{nom}}$, the total variation between the true and model response distributions (P_m and P_s respectively) is given by:

$$d_{\text{TV}}(P_m, P_s) := \frac{1}{2} \sum_{r \in R_q} \big|P_m(r) - P_s(r)\big| \in [0, 1].$$

This can be understood as a special case of the Wasserstein distance, corresponding to optimal transport where the cost is zero when two categories coincide and one otherwise. In this sense, total variation is the natural analogue of Wasserstein for unordered response sets as it quantifies the minimum fraction of probability mass that must be reassigned across categories for the two distributions to align.

C.4 Constructing the Correlation Matrices

In this analysis, we ask: *do simulated responses reproduce the way in which items co-vary across demographic subgroups, as observed in the WVS?* To answer this, we construct and compare correlation matrices derived from the WVS and from each model configuration (OpinionGPT and persona prompting) and do so at two levels of granularity: (i) the ***question–question*** level, which considers correlations between individual survey questions; and (ii) the ***topic–topic*** level, where question are collapsed into the thematic value domains defined in the WVS. The first tests whether models capture fine-grained associations between specific attitudes, while the second evaluates whether they reproduce the broader inter-domain relationships that underpin cultural values theories.

We first aggregate responses by computing the (normalised) ***question mean responses***[8] in the matrix $\text{A}^{\text{WVS}} \in \mathbb{R}^{|Q^{\text{ord}}| \times |S|}$ for the empirical WVS responses, with entries:

$$A^{WVS}_{q,s} = \sum_{r \in R_q} \tilde{r}\, P_s(r \mid q)$$

the mean numerical response to item $q \in Q^{\text{ord}}$ from subgroup $s \in S$ under the response distribution P_s, where $\tilde{r} := \frac{r - \min(R_q)}{\text{diam}(R_q)}$ represents the minmax normalised response value. This is done analogously with the corresponding model distributions P_m.

We then compute the pairwise Pearson correlation coefficients between all questions to produce the ***question-question correlation matrix*** $\text{C} \in \mathbb{R}^{|Q^{\text{ord}}| \times |Q^{\text{ord}}|}$ with elements

$$C_{ij} = \text{corr}\big(\text{A}_{i,\cdot}, \text{A}_{j,\cdot}\big)$$

where $\text{A}_{q,\cdot}$ denotes the vector of mean responses for question $q \in Q^{\text{ord}}$ across all subgroups. To construct the ***topic-topic correlation matrix***, the same procedure applies with the extra step of first aggregating questions to the thematic domains before constructing correlations between topic-level response vectors.

C.5 Bootstrapped Confidence Intervals

To estimate sampling variability for the metrics defined in subsection 3.3, we calculate bootstrapped confidence intervals. Given a model $m \in \{\text{OpinionGPT}, \text{Persona}\}$, which was used to simulate 500 responses to each $q \in Q$ (denoted by X^m), we construct a 95% confidence interval as follows:

For each iteration $b = 1, ..., B = 1000$

1. Randomly resample 500 responses ($\forall q \in Q$) with replacement from the generated responses.

[8] This calculation applies only to ordinal questions $Q^{ord} \subset Q$. Questions with nominal scales constitute less than 10% of the survey, so the correlation structure analysis still covers the vast majority of the questionnaire.

Let the bootstrap sample of responses be denoted by $X^{m,(b)}$

2. Construct the bootstrap response distributions, $P_m^{(b)}$, from $X^{m,(b)}$
3. Compute the correlation matrix from $P_m^{(b)}$ as in subsection 3.3

$$\mathrm{C}^{m,(b)} \in \mathbb{R}^{|Q^{\mathrm{ord}}|\times|Q^{\mathrm{ord}}|}$$

4. Compare $\mathrm{C}^{m,(b)}$ with the true WVS correlation matrix, $\mathrm{C}^{\mathrm{WVS}}$ according to the process outlined in subsection 3.3 to yield the same two similarity metrics

$$c^{(b)} = \mathrm{corr}\big(\mathrm{u}^{(b)}, \mathrm{u}^{\mathrm{WVS}}\big)$$
$$e^{(b)} = \mathrm{RMSE}\big(\mathrm{u}^{(b)}, \mathrm{u}^{\mathrm{WVS}}\big)$$

where u once again denotes the vector obtained by stacking the upper triangle of a correlation matrix C

Finally, the confidence intervals are constructed from the 2.5th and 97.5th percentiles of the bootstrapped estimates $\{c^{(b)}\}_{b=1}^{B}$ and $\{e^{(b)}\}_{b=1}^{B}$, yielding the intervals $[c_{0.025}, c_{0.975}]$ and $[e_{0.025}, e_{0.975}]$.

C.6 Upper and Lower Bound Estimation for Correlation Structure Metrics

To contextualise the evaluation of correlation-structure similarity, we use two complementary procedures: (1) a ***permutation-based null baseline***, corresponding to the absence of cross-item correlation structure; and (2) a ***split-half resampling*** analysis, corresponding to an empirical "best case" where the compared correlation matrices are constructed from subsets of the same distribution. Under their respective scenarios, these provide pessimistic and optimistic estimates on the proposed metrics beyond the trivial bounds of the range of the RMSE and ρ functions ([0, 1] and [-1, 1], respectively).

Permutation-based null baseline. While the unsteered `Phi-3` model serves as a baseline for marginal distribution comparisons, its subgroup-agnostic nature precludes its use for evaluating inter-subgroup correlations. We therefore construct a separate baseline via per-column (i.e., per question or topic) permutation of subgroup means, preserving the empirical distribution of each column's subgroup means while destroying any consistent subgroup ordering shared across questions.

This yields a null distribution for correlation-structure similarity under the hypothesis of no shared inter-question dependence, against which model performance can be interpreted. Model performance can then be interpreted relative to this null: correlation structures that only marginally exceed permutation-based similarity indicate limited recovery of empirical inter-subgroup dependence.

For each iteration $b = 1, ..., B = 1000$

1. For each column, j, of the empirical subgroup mean matrix $\mathrm{A}^{\mathrm{WVS}} \in \mathbb{R}^{|S|\times n}$, independently sample a random permutation of the column values to yield a permuted mean matrix

$$\mathrm{A}^{(b)} \in \mathbb{R}^{|S|\times n}$$

2. Use $\mathrm{A}^{(b)}$ to compute the corresponding correlation matrix

$$\mathrm{C}_{\mathrm{null}}^{(b)} \in \mathbb{R}^{|Q^{\mathrm{ord}}|\times|Q^{\mathrm{ord}}|}$$

3. Calculate the same similarity metrics between correlation matrices, $\mathrm{C}_{\mathrm{null}}^{(b)}$ and $\mathrm{C}^{\mathrm{WVS}}$

$$c_{\mathrm{null}}^{(b)} = \mathrm{corr}\big(\mathrm{u}_{\mathrm{null}}^{(b)}, \mathrm{u}^{\mathrm{WVS}}\big)$$
$$e_{\mathrm{null}}^{(b)} = \mathrm{RMSE}\big(\mathrm{u}_{\mathrm{null}}^{(b)}, \mathrm{u}^{\mathrm{WVS}}\big)$$

We take the mean values, $\bar{c} = \frac{1}{B}\sum_{b=1}^{B} c^{(b)}$ and $\bar{e} = \frac{1}{B}\sum_{b=1}^{B} e^{(b)}$, to define a floor on the correlation and a ceiling on the RMSE, respectively.

The above procedure is outlined for the question-question correlation structures. The same procedure applies for topic-topic level analysis with the additional step of first aggregating responses accordingly.

Split-half resampling analysis. Having established a lower bound via permutation of subgroup means, we next estimate an upper bound on correlation-structure similarity by quantifying the intrinsic stability of the empirical WVS correlation matrix, C^{WVS}, under respondent-level resampling. If C^{WVS} itself is noisy, then no model can reliably exceed that similarity level. To assess this, we apply the procedure outlined in subsection 3.3 as part of a split-half correlation analysis. This then sets a ceiling on the correlation between a true and simulated correlation matrix and floor on the RMSE.

For each iteration $b = 1, ..., B = 1000$

1. Within each of the ten subgroups, randomly split respondents into two halves
2. For each half $h \in \{1, 2\}$, compute the subgroup means over each (ordinal-scaled) question

$$\mathrm{A}^{(b,h)} \in \mathbb{R}^{|S|\times|Q^{\mathrm{ord}}|}$$

Subgroup	True Responses	Opinion GPT	Persona Prompting
Liberal	0.014	0.144	**0.051**
Conservative	0.016	0.070	**0.059**
German	0.024	0.070	**0.054**
American	0.001	0.096	**0.059**
Latin America	0.021	0.107	**0.055**
Middle East	0.028	0.113	**0.055**
Men	0.019	**0.055**	0.058
Women	0.027	0.087	**0.056**
People Over 30	0.021	0.065	**0.062**
Old People	0.025	0.091	**0.061**

Table 8: Share of invalid responses per subgroup for the true responses, OpinionGPT and Persona Prompting.

3. Use the means to construct each correlation matrix

$$C^{(b,h)} \in \mathbb{R}^{|Q^{\text{ord}}|\times|Q^{\text{ord}}|}$$

4. Compute the two metrics, RMSE and Pearson correlation, between the correlation matrices, i.e., between each half of the data

$$c_{1,2}^{(b)} = \text{corr}\big(\text{u}^{(b,1)}, \text{u}^{(b,2)}\big)$$
$$e_{1,2}^{(b)} = \text{RMSE}\big(\text{u}^{(b,1)}, \text{u}^{(b,2)}\big)$$

Again, the resulting means for each metric are used to define the upper bound on model performance. Topic-topic analyses are constructed analogously as before.

D Supporting Results and Analyses

D.1 Invalid Response Rates

Table 8 shows the proportion of invalid responses produced by OpinionGPT and persona promoting for each demographic subgroup compared with the corresponding true non-response rates from WVS Wave 7. The true data contains comparatively fewer non-responses than either modeling approach, with OpinionGPT producing notably more invalid responses than persona prompting. Additionally, OpinionGPT displays much greater differences in invalid response rate across subgroups

Both OpinionGPT and persona prompted models, produced particularly high rates of invalid responses for questions with numeric response options, for which all OpinionGPT modules and persona-prompted models had >10% invalid responses. This was a consistent trend for both model configurations across questions with similar numeric response sets. High rates of invalid responses were also observed on some questions beyond this, although without a clear structural pattern to the questions.

D.2 Modal Collapse

	OpinionGPT	Persona Prompting
Liberal	**6**	17
Conservative	**6**	10
German	**7**	13
American	**0**	11
Latin America	**0**	6
Middle East	**0**	10
Men	**0**	7
Women	**0**	9
People Over 30	**2**	8
Old People	**4**	7

Table 9: No. of questions with only a single response by model

A common consequence of insufficient response diversity is the collapse of the output distribution towards the mode. Table 9 shows the incidence (by number of question) of this ***modal collapse*** for each subgroup. It can be seen that the OpinionGPT modules result in much less modal collapse than persona prompted models; for the unsteered baseline there were 10 questions with modal collapse.

D.3 Effect of Omitting Categorical Questions

In order to construct the correlation matrices (as outlined in Appendix subsection C.4), we first calculate the mean responses. As categorical or nominal response scales lack the notion of a mean response, we leave them out of this analysis. This omits 18 questions of our data subset, or less than 10% of the 193 questions, thus leaving the majority of questions covered by the comparison of correlation structures.

To see the effect this omission would have on marginal dissimilarity scores we can calculate the Wasserstein distances for questions with Likert-type/numerical response scales, shown in Figure 8. Compared to the complete score by subgroup in subsection 3.2, omitting categorical questions only changes dissimilarity scores minimally, ± 0.01 or around $1-4\%$. For the question topics only four[9] have questions with categorical scales with moderate differences for *Perceptions of Security* and *Religious Values* and minimal or no differences for the remaining topics.

[9] *Perceptions of Security*, *Religious Values*, *Political Interest and Political Participation* and *Political Culture and Political Regimes*

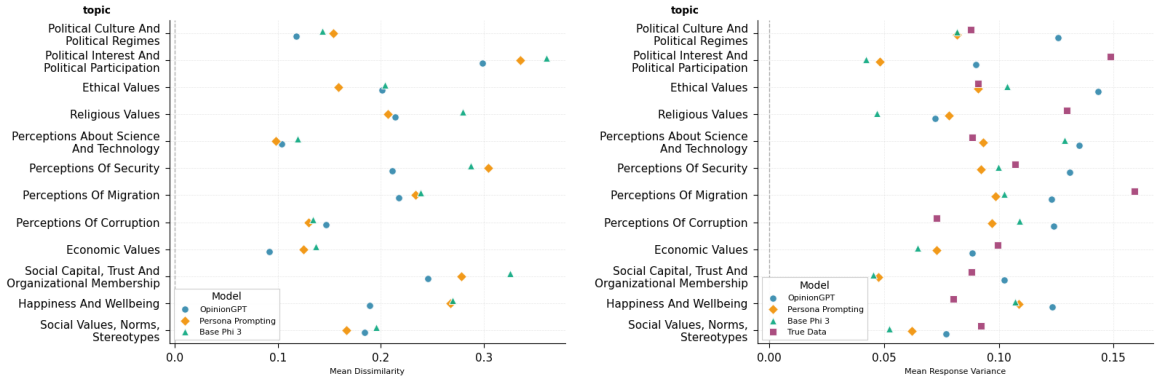

Figure 7: Evaluation metrics by question topic. Left: mean dissimilarity (lower = better). Right: mean response variance (closer to true data = better).

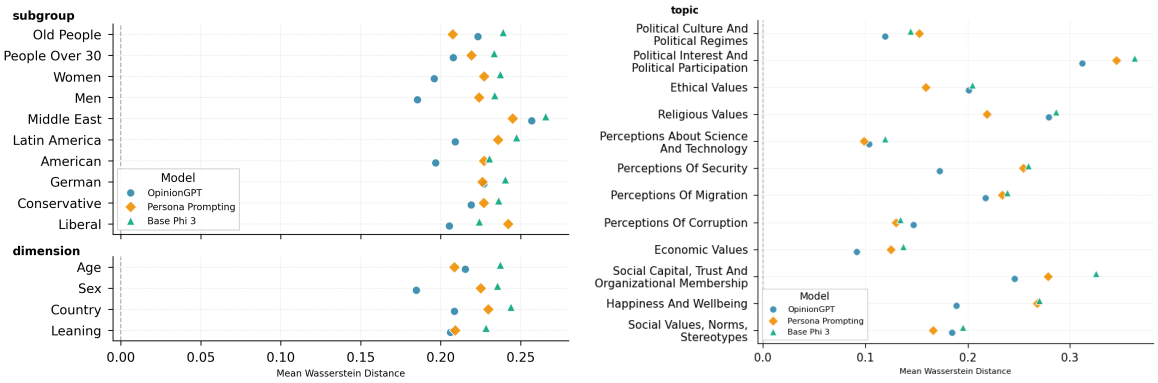

Figure 8: Mean dissimilarity without categorical questions (lower = better). Left: by demographic subgroup and dimension. Right: by question topic

E Ablations

E.1 Temperature Effects

Sampling temperature is a key decoding parameter that controls the entropy of the token distribution and, consequently, the variability of responses generated by large language models

Sampling temperature is a key decoding parameter that controls the entropy of the model's token-level output distribution and, consequently, directly influences the variability of responses generated by large language models (Holtzman et al., 2020). Lower values make the model more deterministic, whereas higher values increase response diversity but may produce erratic or unfaithful outputs. There is no clear, natural value for the temperature that guarantees optimal results. We therefore test the model's marginal dissimilarity and response variance for different temperature settings. The results are in table 10.

The results highlight a trade-off between variance, alignment, modal collapse, and validity. As temperature increased, the variance of both OpinionGPT and persona-prompted responses increased. However, the persona-prompted Phi-3 remained consistently under-dispersed even at high temperature values. Alignment improved at higher temperatures, indicating that moderate stochasticity better captures the population-level patterns. At low temperatures, both models suffered from modal collapse, with several survey items converging onto degenerate distributions, a problem that was substantially alleviated at higher settings (see also Appendix D.2. However, this improvement came

Temp.	OpinionGPT	Persona Prompting
0.6	0.223 *(0.104)*	0.235 *(0.068)*
0.7	0.214 *(0.106)*	0.230 *(0.070)*
0.8	0.205 *(0.108)*	0.224 *(0.073)*
0.9	0.198 *(0.109)*	0.219 *(0.075)*
1.0	0.192 *(0.112)*	0.213 *(0.078)*

Table 10: Average marginal dissimilarity and variance (in parentheses) for persona prompting and OpinionGPT across different temperature settings, averaged across all four sociodemographic dimensions.

alongside a marked increase in invalid responses, particularly pronounced in OpinionGPT relative to persona prompting. The choice of temperature therefore involves a fundamental methodological compromise.

For our evaluation, we choose a temperature of 0.9 in light of improved alignment and low model collapse, while still controlling the invalid response rate to an extent. However, although the exact balance between modal collapse and invalid response rates varies with the choice of temperature, the broader conclusions of this study, concerning the relative dispersion, alignment, and representativeness of the models, remain robust across settings.

E.2 Response Order Effects

To test for response order effects, we randomly flip the response scale for 50% of samples. We then prompt the model with the questions and these original and flipped response scales, normalise the response options to a common [0,1] scale and compare the resulting output distributions by computing per-item means. Table 11 shows these means.

Subgroup	Persona Prompting			OpinionGPT		
	orig	flip	diff	orig	flip	diff
Liberal	0.48	0.26	*0.22*	0.46	0.32	*0.14*
Conservative	0.44	0.27	*0.17*	0.45	0.32	*0.13*
Germany	0.48	0.27	*0.21*	0.50	0.32	*0.17*
America	0.47	0.26	*0.21*	0.49	0.32	*0.17*
Latin America	0.47	0.26	*0.21*	0.51	0.34	*0.16*
Middle East	0.46	0.25	*0.20*	0.52	0.25	*0.28*
Men	0.47	0.26	*0.21*	0.55	0.30	*0.25*
Women	0.48	0.26	*0.22*	0.54	0.26	*0.29*
P. Over 30	0.47	0.27	*0.20*	0.55	0.27	*0.28*
Old People	0.49	0.27	*0.23*	0.56	0.26	*0.30*

Table 11: [0,1]-normalised mean responses for persona prompting and OpinionGPT, with the original (orig) and flipped (flip) order of the answer options and their difference (diff).

The results revealed substantial response order effects across all models, with an average difference of approximately 0.20 between the normalised mean responses under the original and flipped orderings (after remapping to the original scale). In every subgroup, the original ordering produced systematically higher values, consistent with a tendency to favour later options in the list, akin to a recency bias. Although recency effects are recorded more frequently than primacy effects in human surveys, the effects are generally modest (Krosnick and Alwin, 1987; Holbrook et al., 2007). We find some differences between persona prompting and OpinionGPT with persona prompting having more consistent response order effects, but they are not systematic, meaning that these response order effects are likely a result of the base model and not of the finetuned adapters or the persona prompt. One reason for these strong order effects may be the small size of our base model (`Phi-3-mini-Instruct` with 3.5B parameters).

To reduce this response-order biases in our findings, we keep the flipped response order for 50% of the sampled responses to each question and retain the original order for the other 50%. For the evaluation, we reassign the flipped responses to the original scale.