POPULATION FIDELITY: EVALUATING POPULATION REPRESENTATIVENESS IN LLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models (LLMs) show considerable potential in simulating human attitudes and preferences. Prior work finds that LLM-generated responses can compress the range of attitudes found within populations and misrepresent particular subgroups in ways that vary across models and topics. We introduce Population Fidelity, an evaluation framework that distinguishes key conditions required for a set of LLM-generated responses to represent a population. It incorporates three dimensions: group-level accuracy, the amount of between-group variation, and the structure of that variation. We demonstrate the framework's utility in two ways. First, we reproduce a prior study of "machine bias" in LLM survey responses and apply the framework to its models and more recent ones, showing that poor representation reflects not only insufficient between-group variation but also variation assigned to the wrong groups. Second, we evaluate one proposed approach to improving models' population representativeness: cultural fine-tuning. We find that cultural fine-tuning can improve alignment with the survey center without improving the representation of within-population differences, a distinction that unidimensional evaluations overlook. We argue that representing a population requires models to reproduce several features of human attitudinal variation simultaneously. Our framework organizes these features and provides reusable code, data, and trained models for evaluating population fidelity across substantive domains and assessing proposed alignment methods.

1 INTRODUCTION

Large language models (LLMs) show considerable potential in simulating human attitudes, behaviors, and preferences in wide-ranging applied and research contexts (Bail, 2024; Davidson & Karell, 2025; Karetnikov et al., 2026). There is particular interest in using LLMs to generate synthetic survey responses as proxies for human responses. With appropriate prompting and context, the resulting "silicon samples" should reproduce the response distributions of human populations (Argyle et al., 2023; Dillion et al., 2023). This could enable inferences about population-level attitudes on important social, cultural, and political issues, especially as traditional survey methods continue to face threats to their reliability (Keeter et al., 2017; Zhang et al., 2025).

In practice, researchers have raised serious concerns about LLMs' ability to reproduce human response patterns. Several studies, focusing on GPT models, have documented how LLMs reproduce only a limited range of attitudinal variation and misrepresent the positions of particular groups and populations as a whole (Bisbee et al., 2024; Santurkar et al., 2023). Recently, Boelaert et al. (2025) documented a broader pattern of "machine bias," whereby LLMs' responses show unrealistically limited variation across demographic groups and misrepresent human responses in inconsistent ways. Rather than LLMs as a whole being systematically biased toward particular demographic groups, this finding highlights how patterns of (mis-)representation vary across models and topics. This raises the question of how to systematically evaluate such differences across LLMs. This is particularly urgent given that researchers and application developers must choose from among a wide range of models, with distinct representational capabilities, when implementing different types of social simulations.

Model selection is further complicated by proposed methods for improving LLMs' population representativeness. One approach involves fine-tuning language models on empirical data to better re-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

produce observed response distributions, for tasks including missing data imputation in surveys and predicting responses to unasked questions (Kim & Lee, 2026; Suh et al., 2025). CultureLLM, for example, uses cross-national World Values Survey (WVS) responses for fine-tuning culture-specific LLMs that address foundation models' knowledge deficiencies about cultures and languages underrepresented in training corpora (Li et al., 2024). These approaches often target specific forms of alignment, such as matching a model's responses to a *country-level* response distribution for a set of survey questions (Cao et al., 2025). While such fine-tuning can bring model responses closer to those of a target population, it does not necessarily reproduce variation *within* that population and among its demographic subgroups (Wightman et al., 2026). It may even compress or misplace differences among social groups, exacerbating existing sources of misalignment (Adilazuarda et al., 2025). Recent work cautions against inferring alignment from marginal response distributions, which may obscure how models misrepresent populations' underlying attitudinal structures (Williams et al., 2026). We contribute to this line of research with a framework for assessing how well models recover the magnitude and structure of between-group differences within survey items, properties that measures of item-level agreement and cross-item correlation do not directly assess.

We introduce Population Fidelity, an evaluation framework that distinguishes key conditions required for a set of LLM-generated responses to represent a population. It decomposes population representativeness into three dimensions. *Accuracy* measures model responses' correspondence with each population subgroup's observed response distribution. *Adaptability* measures whether the model reproduces the observed amount of variation between subgroups. *Structure* measures whether the same groups are relatively similar or different in generated and observed responses. The Population Fidelity Score (PFS) combines accuracy, adaptability, and structure, providing a summary measure of model representativeness in relation to a given set of survey responses. Considering the overall score alongside its components and *center alignment* (a separate measure of agreement between model and survey centers) demonstrates how LLMs can achieve some aspects of representativeness despite failing to achieve overall fidelity and how better center alignment need not imply better representation of within-population variation.

Figure 1 illustrates why aggregate agreement alone is insufficient to establish Population Fidelity. Each model-generated distribution appears representative: pooled across groups, each reproduces the survey's overall response distribution exactly (the gray area in every panel), and the three have nearly identical accuracy. Yet each misrepresents between-group differences: M1 eliminates them by giving every group the same distribution; M2 exaggerates them, making groups appear farther apart than observed; and M3 reproduces the correct amount of difference but assigns it to the wrong groups. Center alignment and accuracy therefore cannot distinguish these models, whereas adaptability and structure can. M1 and M3 receive $\mathrm{PFS} = 0$ for different reasons, while M2, which preserves which groups are more alike, is penalized mainly for excess variation.

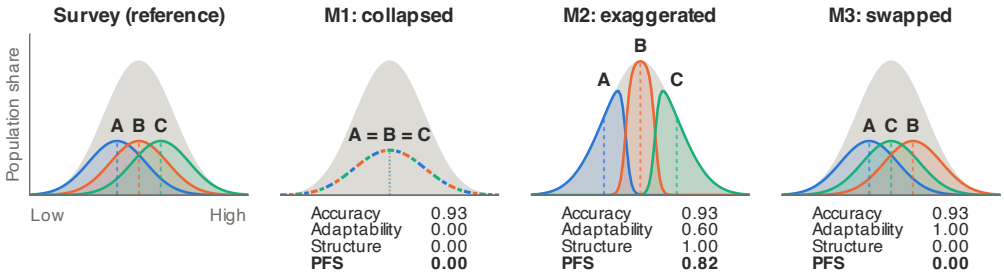


Figure 1: Stylized example with three subpopulations (A–C). Colored curves show each group's share of respondents and sum to the pooled distribution (gray), which is identical in every panel ($S_{\text{center}} = 1$ for all models); dashed lines mark group means. Calculations are in Appendix B.3.

We demonstrate the framework's utility in two ways. First, we reproduce the results of Boelaert et al. (2025) from their original replication materials and apply the framework to both their models and newer ones. Their diagnosis centers on too little variation across demographic groups; incorporating structure shows that this is only part of the problem. Across 164 evaluated combinations of model variant, survey question, and elicitation method, median accuracy is approximately 0.83, whereas median structure is approximately 0.06 and is the limiting PFS component in 159 cases,

and even models that match or exceed the survey's between-group variation rarely reproduce its structure. Second, we evaluate one proposed approach to improving models' population representativeness: cultural fine-tuning. Specifically, we apply CultureLLM-style fine-tuning (Li et al., 2024) and compare German and Mexican adapted models with their corresponding as-released models. CultureLLM-style fine-tuning does not improve PFS on average: mean changes are -0.024 for German and -0.014 for Mexican fine-tuning. Within Mexico, Mexican fine-tuning improves center alignment by 0.038 while target-country PFS changes by -0.010. These results show that fine-tuning can improve alignment with a target survey center without improving the representation of differences within that population, a distinction that unidimensional evaluations overlook.

Overall, we argue that representing a population requires models to reproduce several features of human attitudinal variation simultaneously. Our framework organizes these dimensions and we provide reusable code and data for evaluating population fidelity and alignment methods at https://anonymous.4open.science/r/LLM-population-fidelity-DD50.

2 Related Work

2.1 Misalignment in LLM-simulated human attitudes

Despite its considerable potential (Argyle et al., 2023), recent work has highlighted serious limitations in LLMs' abilities to generate synthetic survey responses that faithfully represent human populations. Although models can reproduce human attitudes at the coarsest levels of analysis, including average responses that are close to survey-based population averages, they fail to reproduce much of the underlying variation within populations, including the true range of positions expressed and their relationships to demographic characteristics (Bisbee et al., 2024). Recent studies document several kinds of misalignment in LLM-generated responses, including bias towards dominant viewpoints and flattening the perspectives of smaller countries and population subgroups (Santurkar et al., 2023; Wang et al., 2025). Boelaert et al. (2025) identify a broader pattern of "machine bias" across a range of models, referring to their tendency to not only misrepresent population groups but to do so inconsistently, favoring different demographic groups in ways that vary arbitrarily across models and topics. These biases constitute potential sources of harm for misrepresented groups (Gallegos et al., 2024), with concerning downstream consequences given the wide-ranging and increasingly consequential scenarios in which LLMs are deployed (Cheng et al., 2023; Lukk, 2026). Because models' representational capabilities vary across models and questions, it is crucial to develop evaluations that assess LLMs' performance on several key aspects of population representativeness.

2.2 Cultural Alignment and Steering Methods

Researchers have proposed several ways to steer models to improve alignment with cultural and population targets. A variety of studies investigate the effects of techniques applied at inference time, most notably demographic persona-based prompting (Argyle et al., 2023; Liu et al., 2024), with evidence that these do not significantly improve model representativeness (Santurkar et al., 2023; Bisbee et al., 2024) and may even worsen certain kinds of social biases (Deshpande et al., 2023; Wan et al., 2023). An alternative approach involves fine-tuning language models on relevant empirical data to better reproduce real-world responses or predict responses on unasked survey questions (Kim & Lee, 2026). CultureLLM uses selected World Values Survey questions as seed data to generate semantically equivalent training examples and fine-tune models aligned with particular cultural perspectives, especially those likely to be underrepresented in model training data (Li et al., 2024). Cao et al. (2025) instead fine-tune models to reproduce country-level response distributions, while Suh et al. (2025) train on distributions for demographic and ideological subpopulations. These promising approaches encode different notions of what successful population adaptation entails, further motivating a framework that distinguishes the aspects of population representation recovered by different alignment methods.

2.3 Evaluating Population Representation

Previous works primarily use single-metric evaluations to compare model responses to those of human survey respondents, including correlation and regression coefficients for individual quantitative

outcomes (Argyle et al., 2023; Bisbee et al., 2024) and distance measures of pairs of opinion distributions (Santurkar et al., 2023; Boelaert et al., 2025). An exception to this is work by Williams et al. (2026), who, in addition to comparing simulated and observed response distributions, measure whether models accurately reproduce how human attitudes are structured *across* survey questions. In contrast to these approaches, we enumerate multiple conditions required for a set of LLM-generated responses to represent a population, with an emphasis on faithful variation among social groups *within* a given question. We develop metrics to assess those conditions and propose a summary Population Fidelity Score to enable comparisons across models and survey items.

3 Machine Bias Design and Reproduction

Our evaluation uses the survey data, demographic cells, prompts, and distance measure of Boelaert et al. (2025). Their design compares LLM responses with three waves of the WVS across happiness, political ideology, religious attendance, and social trust. Respondents are grouped by country, survey wave, sex, age, education, employment, and marital status, yielding 687 demographic cells for three questions and 639 for political ideology. Each cell is represented by an interview-style demographic prompt. Responses are elicited either from renormalized next-token probabilities (NTP) or from sampled full answers (FA). Correspondence between survey and model distributions is measured using normalized Earth Mover's Distance:

$$\mathrm{nEMD}(p, q) = \frac{1}{K-1} \sum_{k=1}^{K-1} \left| \sum_{j=1}^{k} (p_j - q_j) \right|, \tag{1}$$

where p and q are distributions over K ordered response options, so that $\mathrm{nEMD} \in [0, 1]$. This metric supplies the distance used throughout Section 4.

Before applying our framework, we check that our independent implementation reproduces the original results. Passing the archived generations released by Boelaert et al. (2025) through our pipeline reproduces every full-precision checkpoint reported for Mixtral-8x7B, along with the published nEMD distributions, regression analyses, and multidimensional-scaling results. We also regenerate Mixtral-8x7B responses from the public weights using the original prompts. This fresh run is equivalent to the archived one under NTP but not fully under FA; the two runs differ in quantization and inference backend as well as sampling. We retain both as reference series, since the archived run preserves comparability with the published results, and the fresh run provides a reference under our current inference setup. Appendix A reports the reproduction and equivalence tests in full.

4 The Population Fidelity Framework

The analyses of Boelaert et al. (2025) diagnose machine bias at the level of an entire model run. We decompose population fidelity into quantities that can be computed for a complete population or any sufficiently large subset of its subpopulations. The framework distinguishes whether a model approximates particular groups, reproduces the observed amount of variation between groups, and preserves the structure of those differences.

4.1 Unit of Analysis and Notation

Consider a survey question with K ordered response options and a model evaluated on n retained subpopulations, which we call *cells*. Let $i = 1, \ldots, n$ index the cells, with s_i denoting the empirical survey-response distribution and m_i the corresponding model-response distribution. Both are probability vectors over the same response categories. Distances are calculated using Equation 1.

4.2 Population Fidelity Dimensions

Accuracy. Average cell-level error is $E = \frac{1}{n} \sum_{i=1}^{n} \mathrm{nEMD}(s_i, m_i)$. We define the accuracy score as $S_{\mathrm{acc}} = 1 - E$. It measures how closely the model approximates each group's observed response distribution and corresponds to the per-cell distances used in the original Machine Bias study.

Adaptability. Accuracy does not indicate whether the model reproduces variation between social groups. To measure this, we first calculate the median pairwise distances among model and survey cells: $D_{\mathrm{LLM}} = \mathrm{median}_{i<j}\left\{\mathrm{nEMD}(m_i, m_j)\right\}, D_{\mathrm{WVS}} = \mathrm{median}_{i<j}\left\{\mathrm{nEMD}(s_i, s_j)\right\}$. We define the adaptability ratio as $A = \frac{D_{\mathrm{LLM}}}{D_{\mathrm{WVS}}}$. This measures whether the model produces as much between-group variation as the survey. A value of $A = 1$ indicates equal dispersion, while values below and above one indicate compression and amplification, respectively. Because both compression and amplification depart from the survey, we define the symmetric adaptability score as $S_{\mathrm{adapt}} = \min(A, A^{-1})$ for $A > 0$, and set $S_{\mathrm{adapt}} = 0$ when $A = 0$. Halving and doubling the survey's dispersion therefore receive the same score.

Structure. A model can reproduce the overall amount of variation while assigning it to the wrong groups. We compare the complete vectors of pairwise distances using $\rho = \mathrm{corr}\left(\left\{\mathrm{nEMD}(s_i, s_j)\right\}_{i<j}, \left\{\mathrm{nEMD}(m_i, m_j)\right\}_{i<j}\right)$, where corr is Spearman's rank correlation. A value near one indicates that the same pairs of groups are relatively similar or different in the model and survey responses. A value near zero indicates little correspondence between the two structures. The combination $A > 1$ and $\rho \approx 0$, which we call *over-steerability without social fidelity*, occurs when demographic conditioning produces strong differences that do not correspond to those observed among human groups. The pairwise-distance vectors contain $\binom{n}{2}$ entries that share cells and are not independent. We therefore treat ρ as a descriptive effect size rather than interpreting its naive significance test (Mantel, 1967). The structure score is $S_{\mathrm{struct}} = \max(0, \rho)$, so a nonpositive association indicates no recovered population structure. We retain the untransformed ρ for diagnostic analyses, particularly when interpreting values near zero.

4.3 Population Fidelity Score and Center Alignment

Population Fidelity Score (PFS) is the geometric mean of accuracy, adaptability, and structure: $\mathrm{PFS} = \left(S_{\mathrm{acc}} \cdot S_{\mathrm{adapt}} \cdot S_{\mathrm{struct}}\right)^{1/3}$. We use the geometric mean because all three dimensions are required for population fidelity. A synthetic population must approximate the responses of specific groups, reproduce the observed amount of between-group variation, and assign that variation to the corresponding groups. An arithmetic mean lets strong dimensions offset failure on another. The geometric mean instead makes weak components constrain the overall score, as in multidimensional indices such as the Human Development Index (Klugman et al., 2011).

Center alignment (S_{center}). Center alignment is reported alongside the PFS but is not included in it. For a set of retained cells I, define the survey center as $\bar{s}_I = \frac{1}{|I|}\sum_{i \in I} s_i$, and the model center as $\bar{m}_I = \frac{1}{|I|}\sum_{i \in I} m_i$. The distance between them is $C_I = \mathrm{nEMD}(\bar{s}_I, \bar{m}_I)$, and $S_{\mathrm{center}} = 1 - C_I$. We report center alignment over all retained cells and, for cultural fine-tuning analyses, over retained cells in the target country. Because the centers weight retained cells equally, center alignment measures agreement between the model and survey centers rather than a respondent-weighted population mean. A model can match this center while remaining nearly invariant across demographic groups. We therefore report center alignment separately from the PFS since a well-aligned center does not compensate for missing variation or structure.

4.4 Population Fidelity within Demographic Subgroups

Each quantity can be recomputed within a subset of cells. For a demographic group g with cell set $\mathcal{I}_g$, we calculate accuracy, adaptability, structure, center alignment, and PFS using only cells in $\mathcal{I}_g$. We consider levels within seven demographic categories: country, survey wave, sex, age, education, employment, and marital status. For each subgroup, we recompute all metrics using only the cells belonging to that subgroup, rather than averaging population-level scores. This allows the framework to identify population segments for which a model or intervention is more or less faithful.

5 EXPERIMENTAL DESIGN

5.1 MODELS AND CONDITIONS

Appendix B.1 provides an overview of the experimental pipeline. We evaluate six open-weight models spanning approximately 2B–31B parameters: Gemma-4-31B-it, Gemma-4-E4B-it, Qwen3-VL-8B-Thinking, Qwen3-VL-2B-Thinking, Llama-3.2-3B, and Muse-Glimmer-30B. Each is evaluated as released and after cultural fine-tuning toward Germany and Mexico. We also evaluate OpenAI's *GPT-5.6 Terra* (Terra), a proprietary model accessed through its API. Terra is evaluated as released under FA elicitation only because its endpoint does not provide usable NTP at nonzero reasoning effort. Mixtral-8x7B is included through two reproduction references: archived generations from the original Machine Bias study and responses regenerated from the public model weights.

The design contains 84 model–condition–question series: 80 from 20 conditions evaluated on four questions, plus four Terra series. The non-Terra series are evaluated under both NTP and FA, yielding 160 run–mode combinations; Terra is FA-only, for a total of 164 combinations.

5.2 CULTURAL FINE-TUNING

In this study, we follow the approach taken by CultureLLM (Li et al., 2024). For each target culture, the training data combine CultureLLM's selected WVS seed questions with semantically augmented paraphrases and culture-specific target answers derived from aggregate survey responses.

We use CultureLLM's German condition and construct a Mexican condition. CultureLLM's original Spanish partition pools respondents from Mexico and Argentina, whereas our reference population is Mexico. We therefore reconstruct the training records using only Mexican WVS responses. We denote the resulting German and Mexican fine-tuned conditions by $\mathcal{F}_{\text{de}}$ and $\mathcal{F}_{\text{mx}}$, respectively.

Each variant is implemented as a low-rank update to its corresponding as-released model using LoRA with rank $r = 8$ and scaling parameter $\alpha = 16$ (Hu et al., 2021). We use QLoRA for the two largest models (Dettmers et al., 2023). The German and Mexican variants of a model use the same training and inference configuration. None of the four evaluation items—happiness, political ideology, religious attendance, or social trust—appears in the fine-tuning data.

5.3 ELICITATION AND EVALUATION

We use the interview-style prompts (Section 3) under both elicitation modes, NTP and FA; FA responses are sampled at temperature 0.7 and parsed against the offered options. A 32-prompt capacity probe precedes each run and flags an NTP run as non-informative if its mean probability mass on valid answer tokens is below 0.10; none is flagged (Appendix C). A cell enters the analysis only with at least 20 valid FA responses. NTP is deterministic under our decoding configuration and is run once. Because FA samples stochastically, we run each of the 18 locally evaluated conditions three times. For consistency with conditions available for only one FA run, primary results use the first FA run; figures show three-run means and 95% CIs where repeats are available (Appendix E.1).

We compute every metric at three scopes: population-wide, using all retained cells; target-country, using cells from the fine-tuning target country (German for $\mathcal{F}_{\text{de}}$, Mexican for $\mathcal{F}_{\text{mx}}$); and subgroup-specific, recomputing the metric within each level of country, survey wave, sex, age, education, employment, and marital status (Section 4.4). We report fine-tuning effects at both population-wide and target-country scopes.

Each fine-tuned condition is compared only with its corresponding as-released model. Under FA elicitation, the set of retained cells can differ across conditions because a cell must contain at least 20 valid responses for that condition. To ensure that paired comparisons are based on the same population, each fine-tuned condition and its as-released baseline are scored on the intersection of the cells retained by both. Thus, each paired difference reflects changes in model responses on the same cells, rather than differences in cell composition. With six models, four questions, and two elicitation modes, this yields 48 paired comparisons per cultural target.

All primary analyses are conducted separately by question; averages across the four questions are reported only as descriptive summaries. Appendix B.2 details metric implementation and retention; Appendix C provides details on experimental setup, fine-tuning, response coverage, and prompts.

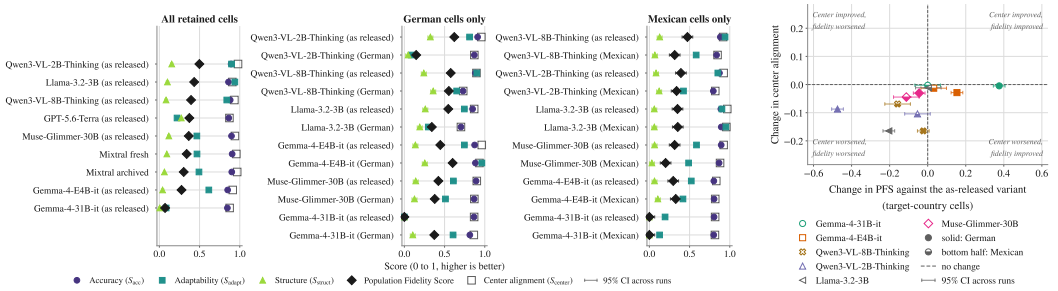
6 RESULTS

6.1 LLMS APPROXIMATE THE SURVEY CENTER BUT MISS GROUP STRUCTURE

Models often achieve high cell-level accuracy but rarely recover the structure of differences among cells. To contextualize accuracy, we use a survey-center baseline that ignores demographic information (Appendix D.1), predicting the equal-cell survey center $\bar{s}_I$ for every cell. This baseline attains $S_{\text{acc}} \in [0.848, 0.915]$ across questions and matches or exceeds model accuracy in 157 of 164 run–mode combinations; only seven combinations outperform it. Thus, much of the observed accuracy reflects proximity to the survey center rather than recovery of demographic differences.

Across the 164 combinations, median $S_{\text{acc}} = 0.832$ whereas median $S_{\text{struct}} = 0.060$ (Appendix E.3, Table 13, default row). S_{struct} is the limiting component in 159 combinations, S_{adapt} in five, and S_{acc} in none. The survey-center baseline itself has $S_{\text{center}} = 1$ but $A = 0$, and therefore $\text{PFS} = 0$, illustrating why center agreement alone is insufficient. Another 30 combinations have $\text{PFS} = 0$ because $\rho \leq 0$; their narrow range, $\rho \in [-0.026, 0]$, indicates absent rather than systematically reversed population structure.

Figure 2a shows this profile for happiness under FA elicitation: accuracy is high for every condition, whereas structure is the weakest component for almost all. The other questions show the same pattern (Appendix D.5). This parallels the machine bias pattern of Boelaert et al. (2025), in which errors are largely explained by each cell's distance from the model's average answer, and a model that barely varies is accurate only insofar as its center sits near the survey cells. Population Fidelity complements this diagnosis by showing that such accuracy can coexist with weak recovery of the observed structure of between-group differences.



(a) Component profile. (b) Fine-tuning shifts.

Figure 2: Population Fidelity for happiness under FA elicitation. **(a)** Components, PFS, and center alignment for the indicated cell sets. **(b)** Fine-tuned minus as-released changes on target-country cells. Markers show three-run FA means using each run's paired support; Table 10, Appendix E.1, uses the stricter all-runs common-cell intersection. All-cell results are in Figure 13, Appendix D.5.

6.2 MORE DEMOGRAPHIC VARIATION DOES NOT IMPLY BETTER GROUP STRUCTURE

Boelaert et al. (2025) identify low adaptability as the core of machine bias. Our ratio A formalizes their comparison of median pairwise distances among model and survey cells (e.g., $0.033/0.115 \approx 0.29$ for Mixtral-8x7B on happiness); structure asks whether the variation a model does produce falls between the right groups. Among the 52 as-released run–mode combinations excluding the Mixtral reference series, 43 compress variation relative to the survey ($A < 1$), whereas nine amplify it ($A > 1$). Compression occurs for every combination on happiness and political ideology, while amplification appears only for religious attendance and social trust (Table 7, Appendix D.3). Yet greater variation rarely corresponds to better group structure: only one of the nine amplifying combinations reaches $\rho > 0.1$ (Figure 12, Appendix D.5, plots A against ρ under NTP). Thus, models can respond differently across demographic profiles without reproducing which groups are relatively

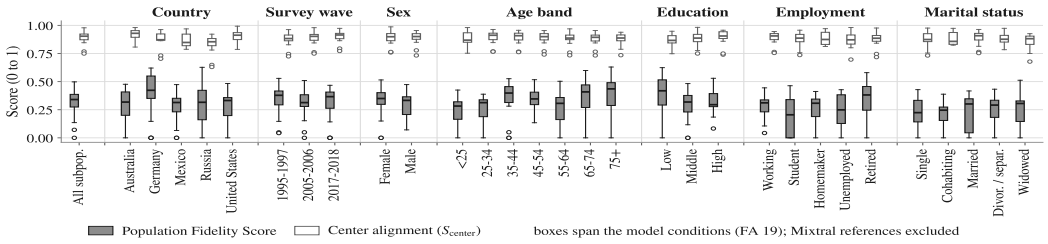


Figure 3: PFS and center alignment by demographic level for happiness under FA elicitation. Boxes summarize 19 model conditions. Results for all questions: Figure 10.

similar or different in the survey. Adaptability and structure therefore capture distinct requirements of population fidelity: matching the amount of variation is not sufficient if that variation is distributed across the wrong groups.

6.3 CultureLLM-Style Fine-Tuning Shows No Consistent Fidelity Gains

In the primary analysis, CultureLLM-style fine-tuning does not improve PFS on average in either scope. Here, ΔPFS denotes fine-tuned minus as-released PFS, so positive values indicate improved population fidelity. Population-wide, effects are mixed: $\mathcal{F}_{\text{de}}$ raises PFS in 21 of 48 comparisons and lowers it in 23 (mean $\Delta\text{PFS} = -0.024$), while $\mathcal{F}_{\text{mx}}$ raises it in 19 and lowers it in 23 (mean -0.014). Target-country means are also negative: -0.021 in Germany and -0.010 in Mexico. Thus, the aggregate result reflects heterogeneous effects rather than a consistent benefit from fine-tuning. Model- and question-specific results are reported in Appendix D.2.

The repeated FA runs support the same conclusion but show greater sampling sensitivity for PFS than center alignment. Across runs, mean ΔPFS on all retained cells and mean ΔS_{center} in both scopes vary by at most 0.005. Within target countries, pooled 95% CI half-widths for individual-comparison ΔPFS are 0.073 in Germany and 0.110 in Mexico, versus 0.005 and 0.003 for ΔS_{center}, respectively (Appendix E.1).

Changes in center alignment do not necessarily track changes in PFS. Figure 2b shows this for happiness within Germany: Gemma-4-31B-it's $\mathcal{F}_{\text{de}}$ increases target-country PFS by 0.371 while barely changing center alignment ($\Delta S_{\text{center}} = -0.004$), whereas Qwen3-VL-8B's $\mathcal{F}_{\text{de}}$ decreases both ($\Delta\text{PFS} = -0.015$, $\Delta S_{\text{center}} = -0.164$). The same appears in the cross-question FA summary. Across models and questions, $\mathcal{F}_{\text{mx}}$ moves responses toward the Mexican survey center ($\Delta S_{\text{center}} = +0.037$) while target-country PFS changes little ($\Delta\text{PFS} = -0.002$); $\mathcal{F}_{\text{de}}$ moves responses away from the German center ($\Delta S_{\text{center}} = -0.087$) while target-country PFS changes by -0.017.

The components clarify this pattern. Within the target countries, fine-tuning reduces adaptability: median S_{adapt} falls from 0.754 to 0.496 in Germany and from 0.616 to 0.421 in Mexico. Structure changes little ($\Delta S_{\text{struct}} = 0.000$ in Germany and -0.004 in Mexico) and remains the limiting PFS component in 44 of 48 $\mathcal{F}_{\text{de}}$ and 47 of 48 $\mathcal{F}_{\text{mx}}$ comparisons. Thus, fine-tuning can shift average responses and alter between-group variation without recovering the observed pattern of group differences. PFS captures these distinctions across fine-tuning objectives. Appendix F illustrates the same diagnostics for CultureLLM-style, distribution-matched, and subgroup-matched fine-tuning on a common backbone. Because these conditions differ in training data, population targets, and evaluation-item overlap, the comparison is diagnostic rather than controlled. On Qwen3-VL-2B-Thinking, the distribution-matched condition retains higher adaptability and structure than the CultureLLM-style variants, whereas the subgroup-matched condition remains structure-limited.

6.4 Center Alignment Can Mask Poor Fidelity within Demographic Subgroups

Within demographic levels, subgroup centers are generally well matched while PFS varies widely, so a matched center is weak evidence of fidelity inside the group. For happiness under FA elicitation, median S_{center} across demographic levels ranges from 0.848 to 0.928 (Figure 3). Across model conditions, PFS and S_{center} are only moderately correlated within demographic levels, with mean Spearman correlations ranging from 0.263 to 0.425 (Appendix D.4).

The imbalance persists across demographic levels, and fine-tuning does not repair it. Across the 4,440 subgroup-level scores (148 run–mode combinations by 30 levels, Mixtral references excluded), S_{struct} is limiting in 4,097 and S_{adapt} in 343, while S_{acc} never is; PFS varies much more across age and employment than across sex: the mean within-condition range is 0.287 for age and 0.27 for employment, versus 0.054 for sex, with these between-level differences driven more by structure than by adaptability (Table 8, Appendix D.4). German and Mexican fine-tuning improve PFS in only 42.5% and 40.8% of subgroup comparisons, respectively; within Germany and Mexico, PFS rises in only 18 of 48 country-level comparisons for each target (Figure 9, Appendix D.4).

6.5 Robustness Across Elicitation, Sampling, and Metric Specifications

The main patterns are broadly consistent across elicitation modes. Among the 20 conditions evaluated under both NTP and FA, PFS rank correlations range from 0.695 for political ideology to 0.880 for religious attendance. Agreement is generally higher for S_{acc} and S_{center}, while S_{adapt} is more sensitive to elicitation, with correlations from 0.502 to 0.913 (Table 12, Appendix E.2). Fine-tuning effects are also directionally similar: NTP and FA agree on the sign of $\Delta\mathrm{PFS}$ in 38 of 48 comparisons, or 35 of 41 after excluding cases with zero change under either mode (Appendix E.2).

The conclusions are also stable across alternative metric specifications. Appendix E.3 varies the dispersion statistic, adaptability transformation, structure correlation and clipping, aggregation rule, and cell weighting. Across these cases, S_{struct} remains the limiting component in at least 156 of 164 combinations, and neither CultureLLM-style target improves population-wide PFS on average under any specification. Replacing the geometric mean with an arithmetic mean changes PFS more because it allows stronger components to offset weak structure. Mean PFS rises from 0.243 to 0.451 and no longer reaches zero, illustrating why the geometric mean is better aligned with our requirement that all three dimensions be jointly satisfied.

7 Discussion, Limitations, and Conclusion

The LLM-generated responses we evaluate approximate survey centers much better than they recover differences among the groups within them. This extends the Machine Bias diagnosis: weak population fidelity is not simply a consequence of too little demographic variation, because some conditions match or amplify the survey's variation while still failing to reproduce which groups are relatively similar or different. Center alignment does not reliably track this failure, either under fine-tuning or across demographic levels. Population Fidelity makes the distinction explicit by separating group-level accuracy, the amount of between-group variation, and its structure. In our experiments, structure is usually the limiting dimension, also under the extra adaptation objectives evaluated in Appendix F.

Our evidence is necessarily limited. We evaluate one survey dataset comprising three WVS waves, five countries, four attitudes, and relatively coarse demographic cells, using English prompts and fine-tuning data. CultureLLM-style fine-tuning is evaluated across six models and two cultural targets, whereas the two additional fine-tuning methods (SimLLCultureDist and SubPOP) are tested on only one backbone; these experiments demonstrate the applicability of PFS across adaptation objectives. Elicitation also matters: adaptability is more sensitive to NTP versus FA than the other components, so compression and amplification should be interpreted within an elicitation mode.

Population Fidelity provides a common framework for asking not only whether an intervention improves alignment with a target population, but which aspects of population representation it changes. Across the evaluations, models rarely recover the observed structure of group differences, while agreement with the survey center often obscures that failure. The additional experiments illustrate how the same framework can distinguish the effects of different fine-tuning objectives. Future work should extend these tests to broader populations, languages, surveys, and model families; scale distribution-based methods beyond a single backbone; and examine richer representations of cultural and individual values, including pluralistic-alignment datasets and established value frameworks.

AI USE STATEMENT

Generative AI tools were used during the preparation of this work to help improve the clarity and organization of the manuscript, to support brainstorming and the organization of ideas, and to assist with software development and debugging. AI-assisted text and code were reviewed and, where relevant, revised and tested by the authors. The authors take full responsibility for the final manuscript, analyses, code, and reported results.

ETHICS STATEMENT

This work evaluates the extent to which LLM-generated survey responses represent patterns observed in human survey data. The analysis uses the World Values Survey through the demographic cells and survey items defined by Boelaert et al. (2025) and does not involve the recruitment of new human participants. Because demographic conditioning and population simulation can reproduce or amplify stereotypes, results for demographic groups should not be interpreted as claims about individual members of those groups. Population Fidelity is intended as an evaluation framework for identifying discrepancies between model-generated and observed population patterns, not as evidence that demographic categories fully characterize the attitudes or behavior of the populations they describe.

REPRODUCIBILITY STATEMENT

We provide the information needed to reproduce the experiments in the main paper and appendices, including the survey and model evaluation procedure, Population Fidelity metrics, model and fine-tuning configurations, elicitation settings, retention criteria, reproduction checks, and sensitivity analyses. The accompanying anonymous repositories contain the code, prompts, configuration files, run manifests, and scripts used to produce the reported results and figures. One repository contains the evaluation of the as-released models and the Machine Bias reproduction,[1] while the second contains the cultural fine-tuning experiments and their evaluation.[2] Appendix A documents the Machine Bias reproduction; Appendix B describes the Population Fidelity framework, metric implementation, and toy example; Appendix C provides the experimental configuration and prompts; Appendix D reports supplementary results; Appendix E presents robustness and sensitivity analyses; and Appendix F evaluates additional adaptation objectives.

ACKNOWLEDGMENTS

Acknowledgments omitted for double-blind review.

REFERENCES

Farid Adilazuarda, Chen Cecilia Liu, Iryna Gurevych, and Alham Fikri Aji. From surveys to narratives: Rethinking cultural value adaptation in llms. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 18063–18090, 2025.

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

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Christopher A. Bail. Can Generative AI improve social science? *Proceedings of the National Academy of Sciences*, 121(21):e2314021121, May 2024. doi: 10.1073/pnas.2314021121.

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

[1] https://anonymous.4open.science/r/LLM-population-fidelity-DD50/README.md

[2] https://anonymous.4open.science/r/CultureFinetuneSteerMLLM-15E1/README.md

Julien Boelaert, Samuel Coavoux, Étienne Ollion, Ivaylo Petev, and Patrick Präg. Machine bias. how do generative language models answer opinion polls? *Sociological Methods & Research*, 54 (3):1156–1196, 2025.

Yong Cao, Haijiang Liu, Arnav Arora, Isabelle Augenstein, Paul Röttger, and Daniel Hershcovich. Specializing large language models to simulate survey response distributions for global populations. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 3141–3154, 2025.

Myra Cheng, Esin Durmus, and Dan Jurafsky. Marked Personas: Using Natural Language Prompts to Measure Stereotypes in Language Models. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1504–1532, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.84.

Thomas Davidson and Daniel Karell. Integrating Generative Artificial Intelligence into Social Science Research: Measurement, Prompting, and Simulation. *Sociological Methods & Research*, 54 (3):775–793, August 2025. ISSN 0049-1241. doi: 10.1177/00491241251339184.

Ameet Deshpande, Vishvak Murahari, Tanmay Rajpurohit, Ashwin Kalyan, and Karthik Narasimhan. Toxicity in chatgpt: Analyzing persona-assigned language models. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 1236–1270, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.88.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms. *Advances in neural information processing systems*, 36:10088–10115, 2023.

Danica Dillion, Niket Tandon, Yuling Gu, and Kurt Gray. Can AI language models replace human participants? *Trends in Cognitive Sciences*, 27(7):597–600, July 2023. ISSN 13646613. doi: 10.1016/j.tics.2023.04.008.

Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed. Bias and Fairness in Large Language Models: A Survey. *Computational Linguistics*, 50(3):1097–1179, September 2024. ISSN 0891-2017. doi: 10.1162/coli_a_00524.

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

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

Nikita Karetnikov, Iyad Rahwan, and Davor Svetinovic. Large language models as human proxies. *Nature Computational Science*, pp. 1–10, September 2026. ISSN 2662-8457. doi: 10.1038/s43588-026-01060-3.

Scott Keeter, Nick Hatley, Courtney Kennedy, and Arnold Lau. What Low Response Rates Mean for Telephone Surveys. Technical report, Pew Research Center, 2017.

Junsol Kim and Byungkyu Lee. AI-Augmented Surveys: Leveraging Large Language Models and Surveys for Opinion Prediction, May 2026.

Jeni Klugman, Francisco Rodríguez, and Hyung-Jin Choi. The hdi 2010: new controversies, old critiques. *The Journal of Economic Inequality*, 9(2):249–288, 2011.

Cheng Li, Mengzhuo Chen, Jindong Wang, Sunayana Sitaram, and Xing Xie. Culturellm: Incorporating cultural differences into large language models. *Advances in Neural Information Processing Systems*, 37:84799–84838, 2024.

Andy Liu, Mona Diab, and Daniel Fried. Evaluating Large Language Model Biases in Persona-Steered Generation. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 9832–9850, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.586.

Martin Lukk. FairFund-Bench: Evaluating Distributive Bias in LLM Resource Allocation. *arXiv preprint arXiv:2607.28934*, 2026. doi: 10.48550/arXiv.2607.28934.

Nathan Mantel. The detection of disease clustering and a generalized regression approach. *Cancer research*, 27(2_Part_1):209–220, 1967.

Meta Superintelligence Labs. Introducing Muse Glimmer: An open agentic model that runs on your device. Meta AI Research blog, 2026. URL https://research.meta.ai/blog/introducing-muse-glimmer-open-agentic-model. Released August 10, 2026.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. Whose opinions do language models reflect? In *International conference on machine learning*, pp. 29971–30004. PMLR, 2023.

Joseph Suh, Erfan Jahanparast, Suhong Moon, Minwoo Kang, and Serina Chang. Language model fine-tuning on scaled survey data for predicting distributions of public opinions. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 21147–21170, 2025.

Gemma Team, Sherif El Abd, Vaibhav Aggarwal, Robin Algayres, Alek Andreev, Olivier Bachem, Ian Ballantyne, Cormac Brick, Victor Cărbune, Michelle Casbon, et al. Gemma 4 technical report. *arXiv preprint arXiv:2607.02770*, 2026.

Yixin Wan, Jieyu Zhao, Aman Chadha, Nanyun Peng, and Kai-Wei Chang. Are Personalized Stochastic Parrots More Dangerous? Evaluating Persona Biases in Dialogue Systems. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 9677–9705, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.648.

Angelina Wang, Jamie Morgenstern, and John P Dickerson. Large language models that replace human participants can harmfully misportray and flatten identity groups. *Nature Machine Intelligence*, 7(3):400–411, 2025.

Maria-Louisa Wightman, Guillaume Bied, and Tijl De Bie. People are not just their countries. disentangling social determinants of llm value alignment across europe. *arXiv preprint arXiv:2608.07367*, 2026.

Tristan Williams, Franziska Weeber, Sebastian Padó, and Alan Akbik. Beyond marginal distributions: A framework to evaluate the representativeness of demographic-aligned llms. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 4806–4826, 2026.

Simone Zhang, Janet Xu, and AJ Alvero. Generative AI Meets Open-Ended Survey Responses: Research Participant Use of AI and Homogenization. *Sociological Methods & Research*, 54(3): 1197–1242, August 2025. ISSN 0049-1241. doi: 10.1177/00491241251327130.

APPENDIX

A MACHINE BIAS DESIGN AND REPRODUCTION DETAILS

This appendix describes the original Machine Bias design, our two-stage reproduction, and the equivalence tests underlying Section 6. Files obtained from the original replication package are recorded in a SHA-256 manifest that is checked before each run. The manifest, code, configurations, and complete reproduction outputs are provided in the accompanying anonymous repository.

A.1 ORIGINAL DESIGN

Boelaert et al. (2025) compare LLM responses with three WVS waves—1995–1997, 2005–2006, and 2017–2018—using four attitudinal items: happiness (4 ordered options), political ideology (10), religious attendance (7), and social trust (2). Within each country–wave, respondents are stratified by sex, age, education, employment, and marital status. Rare demographic profiles are merged until each retained group contains at least 20 respondents. This procedure yields 687 cells for happiness, religious attendance, and social trust and 639 for political ideology, which was not administered in one Russian wave.

Each cell is represented by an interview-style prompt containing its demographic profile. Under next-token probability (NTP) elicitation, the probability mass assigned to valid answer tokens is extracted and renormalized, producing a deterministic response distribution. Under full-answer (FA) elicitation, answers are sampled at temperature 0.7 and parsed using a strict response format. Complete questions and prompt templates appear in Appendix C.4.

Cell-level correspondence is measured using the normalized Earth Mover's Distance in Equation 1. The measure lies in $[0, 1]$ and penalizes disagreement between adjacent response options less than disagreement between distant options. Boelaert et al. (2025) use thresholds of 0.05, 0.10, 0.15, and 0.30 to define quality bands from "very good" to "very bad." They also compare LLM performance with a leave-one-out demographic linear model and random permutations of the survey responses.

The original study identifies machine bias through three findings. First, most model–cell predictions fall into the mediocre-or-worse nEMD bands, performing worse than the demographic baseline and only slightly better than the random baseline. Second, multidimensional scaling shows that model cells occupy a narrower response space than survey cells. For Mixtral-8x7B on happiness, the median pairwise nEMD is 0.033 among model cells and 0.115 among WVS cells. Third, a cell's distance from the model's average response explains much of its prediction error, with R^2 values reaching 0.93–0.95. This factor contributes more than the demographic characteristics in 11 of the 12 model–question combinations.

A.2 ARCHIVED REPRODUCTION

No model inference is performed during the archived reproduction. We pass the released generations through our independently implemented processing and analysis pipeline and compare the resulting counts, distances, regressions, tables, and figures with the published outputs. The full-precision checkpoints available for Mixtral-8x7B on happiness are reproduced exactly in Table 1.

Checkpoint	Value
WVS observations	26,981
Unique NTP profiles	13,904
Happiness cells	687
Mean valid NTP token mass	0.9845
Overall NTP nEMD	0.0308
Overall FA nEMD	0.0354
Median pairwise WVS nEMD	0.1146
Median pairwise NTP nEMD	0.0327

Table 1: Our reproduction matches every full-precision checkpoint reported by Boelaert et al. (2025) for Mixtral-8x7B on happiness. A representative reproduced figure is shown below; complete reproduction outputs are provided in the repository.

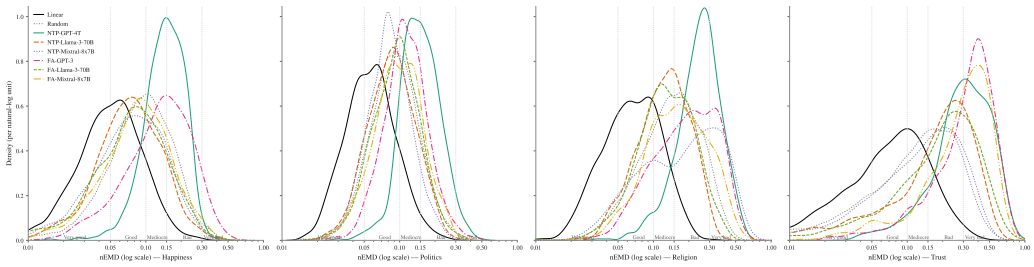


Figure 4: Reproduction of Figure 2 from Boelaert et al. (2025): cell-level nEMD densities for the six published model–elicitation series, shown with the demographic and random baselines and the original response-quality bands.

The pipeline also reproduces Figures 2–4 and 6 from Boelaert et al. (2025), including the nEMD distributions, social-coefficient regressions, multidimensional-scaling geometry, and centroid regressions. We additionally recover the supplementary analyses of prompting strategy, temperature, and backtranslation for all six published model–elicitation series: GPT-4T, Llama-3-70B, and Mixtral-8x7B under NTP, and GPT-3, Llama-3-70B, and Mixtral-8x7B under FA. Figure 4 presents one representative comparison; the complete reproduced outputs are available in the repository.

A.3 Fresh Reproduction

We regenerate Mixtral-8x7B responses using the original single-token, top-1000-logprob NTP procedure and FA sampling at temperature 0.7 with a 12-token generation budget. Each prompt produces an atomic record containing its hash, sampling parameters, backend version, and all FA generation attempts.

The fresh run uses a Q4_K_M GGUF quantization of Mixtral-8x7B-v0.1[3]. Inference uses `llama-cpp-python` 0.3.1 with CUDA and full GPU offloading. The run was executed on a single NVIDIA GeForce RTX 5090 (32 GB).

Archived and fresh runs are compared using difference and equivalence tests with complete country–wave blocks as the resampling unit. The equivalence margin is ± 0.005 nEMD, one tenth of the narrowest quality band used in the original study. Difference-test p-values are obtained from cluster sign-flip permutations and adjusted within each question across the two elicitation modes using the Holm procedure; equivalence tests are reported without multiplicity adjustment.

Under NTP, the archived and fresh runs are equivalent within the prespecified margin on all four questions. For happiness, overall nEMD changes by 2.9% relative to the archived run. Under FA, equivalence is inconclusive for happiness and political ideology. Difference tests detect changes for religious attendance and social trust: mean nEMD changes by $+0.052$ and -0.135, with Holm-adjusted p-values of 0.010 and 0.0002, respectively.

We retain both Mixtral series in subsequent analyses. The archived series preserves comparability with the original study, while the fresh series provides a reproduction reference under the current quantized checkpoint and inference runtime. Because the fresh run differs from the archived run in both quantization and backend as well as stochastic sampling, archived–fresh differences should not be interpreted as estimates of sampling variability alone.

A.4 Additional Reproducibility Controls

For comparability with the original study, each run reproduces its centroid regression on the German cells. A fixed holdout of 20 cells estimates the model center, avoiding use of the same observations to define the center and fit the regression. For the remaining cells, $\log(1+\mathrm{nEMD})$ error is regressed on $\log(1+\mathrm{nEMD})$ distance from that center. Holdout indices reproduce the original R random-

[3]The exact quantized artifact, repository revision, and checksum are pinned in the accompanying code and verified before each run.

number stream using Mersenne–Twister with seed 20,240,110, with stream positions matched by elicitation mode.

FA generation uses a deterministic seed derived separately for each prompt, so interrupting and resuming a run does not change previously written responses. A compatibility flag restores the original unseeded sampling configuration when needed. Unit tests compare the prompt bytes for all four questions and both elicitation modes with those produced by the original implementation.

B Population Fidelity Framework and Metric Details

B.1 Framework Overview

Figure 5 summarizes the Population Fidelity framework. A reference population is partitioned into demographic cells, with each cell represented by an empirical response distribution s_i for a given survey question. The same demographic profiles are then used to elicit responses from the model, producing corresponding distributions m_i.

Population Fidelity compares the survey and model distributions along three dimensions. Accuracy measures correspondence within individual cells, adaptability compares the amount of variation across cells, and structure measures whether the relative differences among cells are preserved. Their geometric mean defines the Population Fidelity Score (PFS). Center alignment compares the equal-cell survey and model centers and is reported separately. All quantities can also be recomputed within demographic subsets.

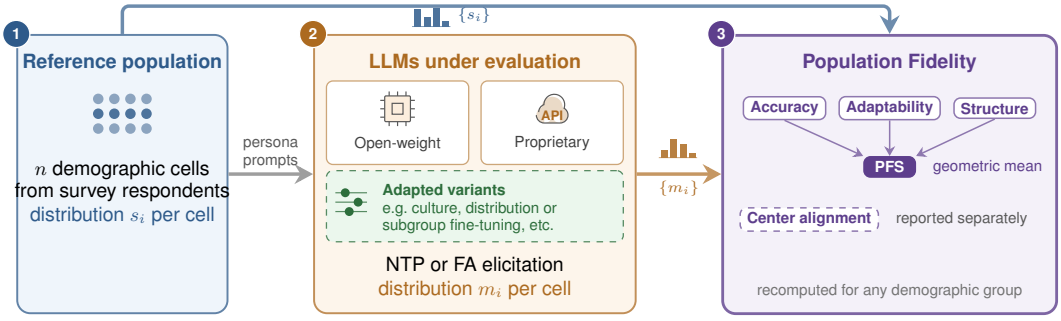


Figure 5: The Population Fidelity framework compares survey and model distributions cell by cell (S_{acc}), in spread (S_{adapt}), and in arrangement (S_{struct}); center alignment (S_{center}) is reported separately. Section 4 defines each. A reference population provides n demographic cells with survey distributions s_i. Model responses to the corresponding demographic profiles produce distributions m_i. Accuracy, adaptability, and structure compare the two sets of distributions and combine into the PFS by geometric mean; center alignment is reported separately. The metrics can also be recomputed within demographic subsets.

B.2 Metric Implementation and Validation

This appendix gives the implementation behind every number in Section 6: distances, cell retention, degenerate cases, cross-question aggregation, and the validation checks every run passes.

Distances. Equation 1 is implemented as the city-block norm of the difference between the cumulative survey and model distributions, divided by $K - 1$. Pairwise distances are stored as condensed vectors using the same cell ordering for survey and model responses. Although nEMD is normalized to $[0, 1]$, differences in response scales and item meaning make averages across questions difficult to interpret. We therefore compute and interpret all distance-based quantities separately by question.

Cell-level model distributions. Multiple elicited profiles can map to the same demographic cell. Under NTP, let $\mathcal{R}_i$ denote the profiles assigned to cell i and q_{ir} the response distribution elicited for profile $r \in \mathcal{R}_i$, and n_{ir} the number of WVS respondents in cell i with that profile. Following

Boelaert et al. (2025), we average over respondents

$$m_i = \sum_{r \in \mathcal{R}_i} w_{ir} q_{ir}, \qquad w_{ir} = \frac{n_{ir}}{\sum_{r' \in \mathcal{R}_i} q_{ir'}}.$$

Under NTP, q_{ir} is the renormalized valid-token probability vector; under FA, one answer is sampled per respondent and m_i is the empirical distribution of valid sampled responses assigned to cell i.

Retention and evaluation support. The survey contains 687 demographic cells for happiness, religious attendance, and social trust, and 639 for political ideology. A cell is scored only when its survey-response proportions are complete. Under FA, the cell must also contain at least 20 valid model responses. Across runs, the retained cell count ranges from 606 to 687.

Population-wide analyses use all retained cells. Target-country analyses use German cells for German fine-tuning and Mexican cells for Mexican fine-tuning. The complete German subset contains 144 cells and 10,296 cell pairs, while the complete Mexican subset contains 131 cells and 8,515 pairs. The demographic-subgroup analysis recomputes all metrics within levels of country, survey wave, sex, age, education, employment, and marital status. Across all 164 combinations, this yields 4,920 subgroup-level scores; the main subgroup analysis excludes the two Mixtral reference series, leaving 4,440 scores across 148 combinations.

For each paired fine-tuning comparison, the fine-tuned and as-released conditions are recomputed on their common set of retained cells. This prevents differences in response coverage from being interpreted as changes in population fidelity. The common set is the intersection of the cells retained in both conditions and is used for both population-level and subgroup-level comparisons. Across the primary paired comparisons, these intersections contain between 617 and 687 cells. Repeated FA runs reuse the same model weights, prompts, decoding configuration, and evaluation procedure, differing only in the seed stream (Section 5). Retained cell sets are highly stable across repeated runs; Appendix E.1 reports the corresponding retention and run-to-run variability analyses.

Under NTP, prompts assigning zero probability mass to all valid answer tokens are treated as failed records rather than renormalized. No such failure affects the retained cells in the reported analyses. For conditions evaluated under both NTP and FA, the two elicitation modes use the same retained cell set.

Undefined and degenerate cases. Adaptability requires at least two cells and structure at least three. Below these thresholds, the corresponding quantities remain undefined. If the model's pairwise-distance vector has zero variance, ρ is undefined; the model is classified as flat and assigned a structure score of zero. No population-wide scored run–mode combination meets this condition; two country-level subgroup estimates do and are assigned $S_{\text{struct}} = 0$ under this rule (Table 8). The nearest case is Qwen3-VL-2B-Thinking under German fine-tuning on social trust with FA elicitation. Its median pairwise distance is zero, so $A = 0$, but its distance vector has sufficient variation for ρ to remain defined. The PFS equals zero when any component is zero and remains undefined when a required component is undefined.

Cross-question aggregation. The cross-question PFS is the geometric mean of the four per-question PFS values. Component summaries are arithmetic means of their per-question values, and the limiting component is identified from these averaged scores. The cross-question PFS is therefore not recomputed from the averaged components. Because a zero PFS on any question produces a zero cross-question geometric mean, per-question values remain the primary results.

Implementation checks. Before producing tables or figures, the implementation verifies that:

1. every recomputed cell-level nEMD matches the corresponding run-level distance table row by row;
2. pooled quantities match independently calculated population summaries;
3. D_{WVS} is identical across model conditions evaluated on the same question, elicitation mode, and retained cell set;
4. all transformed component and PFS values lie within $[0, 1]$;

5. group-level cell counts do not exceed those of their pooled parent populations; and
6. the survey and model pairwise-distance vectors contain the same cells in the same order.

All 164 run–mode combinations pass these checks.

Center alignment in the Population Fidelity Framework gives equal weight to retained cells. For comparability with the original study, the reproduction output also preserves its respondent-weighted center. Both quantities are retained and labeled according to their weighting scheme.

B.3 Stylized Example Calculations

Figure 1 uses three subpopulations, A–C. Their survey distributions are normal with means 4, 5, and 6 and standard deviation 1.3, on a 0–10 attitude scale with 101 response options $(0, 0.1, \ldots, 10)$. The pooled survey distribution $\bar{s}$ is their average.

M1 assigns $\bar{s}$ to every group. M2 re-sorts the same respondents into three narrower groups, each still one third of respondents, with means 3.34, 5.00, and 6.66 and standard deviations 0.81, 0.44, and 0.81. M3 exchanges the survey distributions of B and C. Because each model only reassigns the pooled respondents across groups, each averages to $\bar{s}$, so $S_{\text{center}} = 1$ for all three models. Values are rounded to two decimals, derived quantities use unrounded values, and the exact construction is included in the released code.

The cell-level errors for M1, M2, and M3 are, respectively,

$$[0.10, 0.02, 0.10], \qquad [0.07, 0.07, 0.07], \qquad [0.00, 0.10, 0.10].$$

Each therefore has mean error $E \approx 0.07$ and accuracy $S_{\text{acc}} \approx 0.93$.

Using pair order (AB, AC, BC), the pairwise-distance vectors are

$$d_s = [0.10, 0.20, 0.10], \qquad d_{\text{M1}} = [0, 0, 0],$$

$$d_{\text{M2}} = [0.17, 0.33, 0.17], \qquad d_{\text{M3}} = [0.20, 0.10, 0.10].$$

Their median pairwise distances are 0.10 for the survey, 0 for M1, 0.17 for M2, and 0.10 for M3. The adaptability ratios are therefore $A = 0$, 1.67, and 1.00, and the adaptability scores are 0, 0.60, and 1.00.

M1's pairwise distances are all equal, so ρ is undefined and M1 is assigned $S_{\text{struct}} = 0$ (Appendix B.2). Treating distances that are equal by symmetry as tied, the Spearman correlations with the survey distance vector are 1.00 for M2 and -0.50 for M3. After applying $S_{\text{struct}} = \max(0, \rho)$, the structure scores are 0, 1.00, and 0. Finally,

$$\text{PFS}_{\text{M1}} = 0, \qquad \text{PFS}_{\text{M2}} \approx 0.82, \qquad \text{PFS}_{\text{M3}} = 0.$$

Thus, the three models reproduce the pooled survey distribution exactly and have nearly identical accuracy, but differ in the amount and structure of their between-group variation. M1 corresponds to the survey-center baseline (Appendix D.1).

C Experimental Configuration and Prompts

C.1 Models and Fine-Tuning Configuration

Model inventory. Table 2 summarizes the six open-weight models used in the main evaluation. Each is evaluated as released and after German and Mexican fine-tuning. Within each model, the as-released and fine-tuned conditions use the same tokenizer, numerical configuration, and inference settings. Exact repository revisions are recorded in the released run manifests.

Mixtral-8x7B is retained separately as a reproduction reference. The archived condition uses generations released with the original Machine Bias replication package, while the fresh condition regenerates responses from Mixtral-8x7B-v0.1 using the original prompts and elicitation procedures. All local open-weight training and inference experiments were run on a single NVIDIA GeForce RTX 5090 (32 GB). The backend used for the fresh Mixtral reproduction is described in Appendix A.

LLM	Precision	Quantization	Adapted modules	Transformers (inference)
Gemma-4-31B-it (Team et al., 2026)	bf16	NF4	230	5.8.1
Gemma-4-E4B-it (Team et al., 2026)	bf16	–	132	5.8.1
Qwen3-VL-8B-Thinking (Bai et al., 2025)	bf16	–	144	5.8.1
Qwen3-VL-2B-Thinking (Bai et al., 2025)	bf16	–	112	5.8.1
Llama-3.2-3B (Grattafiori et al., 2024)	bf16	–	112	5.8.1
Muse-Glimmer-30B (Meta Superintelligence Labs, 2026)	bf16	NF4	208	5.15.0

Table 2: Open-weight model inventory. NF4 is 4-bit NormalFloat quantization implemented with `bitsandbytes` 0.49.2. Adapted-modules column reports the number of modules receiving LoRA updates. Every model runs in bfloat16 compute with `PEFT` 0.19.1 and PyTorch 2.12.0 under Python 3.11.

Fine-tuning configuration. For each target culture, the training data combine the WVS seed questions selected by CultureLLM with semantically augmented paraphrases. Each record is paired with a culture-specific target answer derived from aggregate survey responses. The German condition uses CultureLLM's German partition. For the Mexican condition, we reconstruct the records using only Mexican respondents rather than CultureLLM's original Spanish partition, which pools respondents from Mexico and Argentina.

Fine-tuning uses LoRA with rank $r = 8$, scaling parameter $\alpha = 16$, and dropout 0.05. Adapters are applied to the query, key, value, and output projections of the text-attention blocks (`q_proj`, `k_proj`, `v_proj`, and `o_proj`); vision and audio components are excluded. QLoRA with NF4 quantization is used for Gemma-4-31B-it and Muse-Glimmer-30B. Within each model family, the German and Mexican variants use the same training configuration. None of the four evaluation questions appears in the fine-tuning data.

Training uses supervised fine-tuning on chat-formatted records with AdamW, no weight decay, a cosine learning-rate schedule, an effective batch size of 16, gradient clipping at 0.3, a maximum sequence length of 512 tokens, and seed 42. Warmup lasts 50 optimizer steps, or 100 steps for the QLoRA models. Training is allowed to continue for at most 250 epochs and is stopped early based on validation loss after a minimum of 20 epochs, with patience of 15 evaluations. Validation is performed every 100 optimizer steps, and the checkpoint with the lowest validation loss is retained. Table 3 reports the learning rate, batch size, run length and held-out accuracy of each model.

				German		Mexican	
LLM	Adapter	LR	Batch	Epochs (steps)	Val. acc.	Epochs (steps)	Val. acc.
Gemma-4-31B-it	QLoRA, NF4	10^{-4}	1×16	33 (3,900)	0.97	45 (2,700)	0.88
Gemma-4-E4B-it	LoRA, bf16	2×10^{-4}	2×8	37 (4,400)	0.97	45 (2,700)	0.87
Qwen3-VL-8B-Thinking	LoRA, bf16	2×10^{-4}	2×8	58 (6,900)	0.97	45 (2,700)	0.87
Qwen3-VL-2B-Thinking	LoRA, bf16	2×10^{-4}	4×4	58 (6,900)	0.97	45 (2,700)	0.86
Llama-3.2-3B	LoRA, bf16	2×10^{-4}	4×4	80 (9,500)	0.88	45 (2,700)	0.77
Muse-Glimmer-30B	QLoRA, NF4	10^{-4}	1×16	43 (5,100)	0.98	45 (2,700)	0.89

Table 3: Fine-tuning configuration and run length by model and target culture. Batch reports the per-device batch size $\times$ gradient-accumulation steps, yielding an effective batch size of 16. Parentheses report optimizer steps at termination. Val. acc. is the held-out token accuracy of the retained checkpoint.

The German dataset contains 2,100 question–answer pairs, split into 1,890 training and 210 validation records. The Mexican dataset is constructed from the same 50 WVS seed questions and 1,000 augmented paraphrases, using aggregate responses from Mexican respondents only. It contains 1,050 pairs, split into 945 training and 105 validation records.

All fine-tuning records are in English: each contains a system prompt naming the target country, a WVS-derived question with numbered response options, and the corresponding target answer. The experiment therefore evaluates cultural adaptation within a common language rather than cross-lingual transfer. Training was performed on the hardware described in Appendix A.

C.2 Response Classification, Capacity, and Coverage

Response classification. Each FA generation is classified as *valid* if it maps unambiguously to one offered response option, *invalid* if it does not satisfy the required format, and *failed* if no generation is completed. Invalid and failed outputs are retained for audit. For NTP, valid-token probability mass is measured before renormalization. A response assigning no probability mass to any valid token is recorded as degenerate rather than silently renormalized. Coverage records and the original generated outputs are included in the repository.

Capacity and response coverage. Each run receives a 32-prompt preflight probe. Under NTP, a run is flagged as non-informative if the mean probability mass assigned to valid answer tokens is below 0.10. No NTP run in the reported evaluation falls below this threshold.

Under FA, we do not apply a separate run-level threshold. Distances are computed only for cells containing at least 20 valid responses, and dropped cells are recorded for each run. Across the evaluated FA conditions, the lowest overall valid-answer rate is 0.926.

Fine-tuning health. Before elicitation, every adapter used in the reported experiments is checked using the mean norm of the weight update applied to the attention projections and held-out token accuracy against a 50% usability threshold. All evaluated adapters pass this check, with update norms between 0.2 and 5.4 and held-out accuracy between 77% and 98%. No reported Population Fidelity result therefore comes from a checkpoint classified as a training failure.

C.3 API-Served Model

API-served model. GPT-5.6 Terra is evaluated through a commercial chat endpoint using the same survey prompts as the open-weight models. Requests use the model identifier `gpt-5.6-terra` and were issued in August 2026. Reasoning effort is fixed at *medium*, FA responses are sampled at temperature 0.7, and the maximum output length is 300 tokens. A seed of 20240110 is passed with each request as a best-effort parameter.

At nonzero reasoning effort, the endpoint does not provide usable next-token probabilities for our NTP procedure. In a single-token probe requesting the top 20 log probabilities, none of the valid survey-answer tokens appeared among the returned candidates. Terra is therefore evaluated under FA only.

Each request records the prompt, requested sampling parameters, parameters applied by the endpoint, and returned output. Prompts are submitted as a single user message rather than as raw completions.

Computational scale. The experiment requires approximately $\approx$ 7.9 million model queries, excluding the archived Mixtral generations. Each open-weight NTP run evaluates 13,904 unique demographic profiles, while each FA run evaluates 26,981 respondent-level prompts.

C.4 Questions and Prompt Construction

Every response analysed in this paper comes from the four survey items and the prompt template below, used as described in Section 3 and 5.3.

Prompts for the four items in Table 4 follow the interview-style template used by Boelaert et al. (2025). As illustrated in Figure 6, the interviewer asks a sequence of sociodemographic questions and the interviewee supplies the values defining the relevant cell. The model then completes the answer to the final attitudinal question. Categorical demographic questions use lettered options and expect a single capital letter. Under NTP elicitation, the probability mass assigned to the valid first-token options is extracted and renormalized.

Political ideology follows the original study's mode-specific encoding. NTP uses a 0–9 scale so that each option corresponds to a single digit token, whereas FA uses the survey's 1–10 scale. The two encodings are aligned by ordinal position before the response distributions are compared. The NTP prompt for political ideology includes a trailing space after `Answer:` to preserve the original tokenization. Political ideology is unavailable for 48 cells from a survey wave in which the item was

Variable	Topic	Options
d_happy	Happiness	4
d_polpos	Political ideology	10
d_religiousp	Religious attendance	7
d_trust	Social trust	2

Table 4: Attitudinal items used in the evaluation.

```
Question: What year is it?
Answer: {{survey year}}
Question: What country are we in? Is it 'A. Australia' or 'B. Germany' or 'C. Mexico' or 'D. Russia' or 'E.
United States'?
Answer: {{lettered country}}
Question: How old are you?
Answer: {{age}}
Question: What is your gender? Is it 'A. Male' or 'B. Female'?
Answer: {{lettered sex}}
Question: What is the highest educational level that you have attained? Is it 'A. Low' or 'B. Middle' or 'C.
High'?
Answer: {{lettered education}}
Question: Are you employed now or not? Are you 'A. Working' or 'B. Retired' or 'C. Homemaker' or 'D. Student'
or 'E. Unemployed'?
Answer: {{lettered employment}}
Question: What is your marital status? Are you 'A. Married' or 'B. Cohabiting' or 'C. Divorced or separated'
or 'D. Widowed' or 'E. Single'?
Answer: {{lettered marital status}}
Question: {{attitudinal item}}
Answer:
```

The attitudinal item, per question:

d_happy: Taking all things together, how happy are you? Are you 'A. Very happy' or 'B. Quite happy' or 'C. Not very happy' or 'D. Not at all happy'?

d_polpos: In political matters, people talk of "the left" and "the right." How would you place your views on a scale of {{low}} to {{high}}, where {{low}} is left and {{high}} is right, generally speaking? ({{low}}–{{high}} is 0–9 for NTP and 1–10 for FA)

d_religiousp: Apart from weddings, funerals and christenings, about how often do you attend religious services these days? Is it 'A. More than once a week' or 'B. Once a week' or 'C. Once a month' or 'D. Only on special holy days' or 'E. Once a year' or 'F. Less often' or 'G. Never, practically never'?

d_trust: Generally speaking, would you say that most people can be trusted or that you need to be very careful in dealing with people? Is it 'A. Most people can be trusted' or 'B. Need to be very careful'?

Figure 6: Interview-style prompt template. Placeholders are replaced with the demographic values defining each cell; survey year and age are inserted as integers, while categorical attributes use the corresponding lettered options. Demographic items absent from a merged profile are omitted. The final question is replaced by one of the four attitudinal items in Table 4, and the prompt ends at `Answer:` for the model to complete.

not administered; these cells are treated as having undefined targets and excluded from analyses of that question. Unit tests verify the prompt bytes for all four questions and both elicitation modes against the original implementation.

D Supplementary Results

D.1 Survey-Center Baseline for Accuracy

Accuracy is bounded, and on the nEMD scale a prediction that ignores demographic information can already attain a high score. We therefore compare model accuracy with such a baseline. For the retained cells I of a run, we take the equal-cell survey center $\bar{s}_I$ defined in Section 4.3 and predict it for every cell, $m_i^{(0)} = \bar{s}_I$ for all $i \in I$. Its accuracy is

$$S_{\text{acc}}^{(0)} = 1 - \frac{1}{|I|} \sum_{i \in I} \text{nEMD}\left(s_i, \bar{s}_I\right),$$

which is the accuracy attained without using demographic information. We also express model performance relative to this baseline using the skill-score form

$$S_{\text{acc}}^{\text{above baseline}} = \frac{S_{\text{acc}} - S_{\text{acc}}^{(0)}}{1 - S_{\text{acc}}^{(0)}}.$$

This score is zero at the survey-center baseline, one for perfect cell-level prediction, positive when a model exceeds the baseline, and negative when it performs worse.

We score the baseline using the same implementation as the model conditions, so its other components follow directly from the definitions in Section 4. Its center alignment is one by construction. Because all predicted cells are identical, all pairwise model distances are zero, giving $A = 0$ and $S_{\text{adapt}} = 0$; the flatness rule of Appendix B.2 sets $S_{\text{struct}} = 0$, and therefore $\text{PFS} = 0$. Because $\bar{s}_I$ includes each cell's own survey distribution, we also compute a leave-one-out version in which cell i is excluded from the center used to predict it. This changes $S_{\text{acc}}^{(0)}$ by less than 0.0003 at the population level and by at most 0.005 within demographic levels, indicating that the baseline's performance is not driven by including each cell in its own reference center.

Table 5 compares the baseline with the models, with each run evaluated on its own retained cells. The survey-center baseline attains $S_{\text{acc}}^{(0)}$ between 0.848 for religious attendance and 0.915 for political ideology and exceeds the median model accuracy for every question and elicitation mode. Only 7 of the 164 combinations exceed their corresponding baseline, by at most 0.036 in accuracy: on happiness, both Mixtral runs under NTP, the regenerated Mixtral run under FA, and the as-released and Mexican Muse-Glimmer-30B conditions under NTP; on religious attendance, the Mexican Muse-Glimmer-30B condition under NTP; and on social trust, as-released Qwen3-VL-8B under NTP. The median above-baseline score is -0.56, and 52 combinations fall below -1, meaning that their mean cell-level error is more than twice that of the corresponding survey-center baseline.

Question	Mode	Cells	$S_{\text{acc}}^{(0)}$	Model S_{acc}	$S_{\text{acc}}^{\text{above baseline}}$	Above baseline
Happiness	FA	678–687	0.901–0.902	0.860 [0.749, 0.906]	-0.41 [-1.53, 0.05]	1 / 21
	NTP	678–687	0.901–0.902	0.871 [0.769, 0.911]	-0.30 [-1.33, 0.11]	4 / 20
Political ideology	FA	606–639	0.913–0.915	0.875 [0.817, 0.905]	-0.45 [-1.11, -0.09]	0 / 21
	NTP	626–639	0.913–0.914	0.892 [0.848, 0.910]	-0.25 [-0.77, -0.03]	0 / 20
Religious attendance	FA	677–687	0.848	0.707 [0.441, 0.833]	-0.92 [-2.67, -0.10]	0 / 21
	NTP	677–687	0.848	0.754 [0.480, 0.884]	-0.61 [-2.41, 0.24]	1 / 20
Social trust	FA	630–687	0.867–0.868	0.683 [0.350, 0.855]	-1.39 [-3.90, -0.09]	0 / 21
	NTP	630–687	0.867–0.868	0.703 [0.359, 0.883]	-1.24 [-3.84, 0.11]	1 / 20
All questions	both	606–687	0.848–0.915	0.832 [0.350, 0.911]	-0.56 [-3.90, 0.24]	7 / 164
Demographic levels	both	24–377	0.817–0.943	0.824 [0.185, 0.952]	-0.72 [-11.58, 0.40]	157 / 4,440
Target countries	both	130–144	0.875–0.943	0.816 [0.384, 0.942]	-1.24 [-4.89, 0.01]	1 / 96

Table 5: Model accuracy relative to the survey-center baseline. $S_{\text{acc}}^{(0)}$ is obtained by predicting the equal-cell survey center for every retained cell, so its range reflects differences in retained cells across runs. Model S_{acc} and the above-baseline score report the median across combinations, with the minimum and maximum in brackets; the final column counts combinations whose accuracy exceeds their corresponding baseline. For demographic-level results, the baseline is recomputed within each level. Target-country results compare German fine-tuned conditions within Germany and Mexican fine-tuned conditions within Mexico. The baseline has $\text{PFS} = 0$ in every case.

The same pattern holds within demographic levels, where each level is compared with the center of its own retained cells. Across the 4,440 subgroup-level scores of Section 6, models exceed their level-specific survey-center baseline in 157 cases. Within the target countries, German fine-tuned conditions never exceed the German survey-center baseline (0 of 48 comparisons), while Mexican fine-tuned conditions exceed the Mexican baseline once (1 of 48).

This baseline clarifies why high accuracy does not imply population fidelity. Although models exceed the baseline's structure score of zero in 134 of the 164 combinations, they exceed its accuracy in only seven. Much of their absolute accuracy therefore reflects proximity to the survey center, consistent with the machine-bias pattern of Boelaert et al. (2025). The baseline also illustrates the behavior that PFS is designed to reject: a prediction that outperforms nearly all evaluated combinations on accuracy and matches the survey center exactly receives $\text{PFS} = 0$ because it contains no variation between groups.

We report the above-baseline score as a diagnostic and do not include it in PFS. Within each question, the baseline varies by at most 0.0015 across runs, so the above-baseline score produces almost the same ordering as S_{acc}, with Spearman correlations of at least 0.998. It therefore changes the reference point while barely affecting model rankings. Unlike the PFS components, which lie in

$[0, 1]$, the above-baseline score can take large negative values when model error substantially exceeds baseline error; in our population-level results it reaches -3.90.

D.2 Model- and Question-Specific Results

Under FA elicitation, fine-tuned conditions attain the highest PFS on three of the four questions, all with Muse-Glimmer-30B: its German condition leads on political ideology and social trust, while its Mexican condition ties GPT-5.6 Terra on religious attendance after rounding. Happiness is led by as-released Qwen3-VL-2B (Table 6).

LLM	Condition	Happiness	Politics	Religious	Trust
Mixtral (archived)	as released	0.303	0.335	0.142	0.182
Mixtral (fresh)	as released	0.340	0.360	0.073	0.247
GPT-3 (archived)	as released	0.245	0.317	0.000	0.221
GPT-5.6 Terra	as released	0.373	0.270	**0.726**	0.416
Gemma-4-31B-it	as released	0.072	0.381	0.128	0.300
	$\mathcal{F}_{\text{de}}$	0.329	0.303	0.000	0.000
	$\mathcal{F}_{\text{mx}}$	0.000	0.410	0.000	0.301
Gemma-4-E4B-it	as released	0.284	0.162	0.000	0.000
	$\mathcal{F}_{\text{de}}$	0.409	0.253	0.119	0.000
	$\mathcal{F}_{\text{mx}}$	0.248	0.315	0.117	0.000
Qwen3-VL-8B-Thinking	as released	0.413	0.274	0.000	0.382
	$\mathcal{F}_{\text{de}}$	0.420	0.227	0.000	0.117
	$\mathcal{F}_{\text{mx}}$	0.294	0.277	0.000	0.178
Qwen3-VL-2B-Thinking	as released	**0.488**	0.285	0.000	0.233
	$\mathcal{F}_{\text{de}}$	0.141	0.194	0.225	0.000
	$\mathcal{F}_{\text{mx}}$	0.338	0.226	0.000	0.000
Llama-3.2-3B	as released	0.457	0.330	0.000	0.284
	$\mathcal{F}_{\text{de}}$	0.273	0.322	0.112	0.119
	$\mathcal{F}_{\text{mx}}$	0.385	0.289	0.187	0.300
Muse-Glimmer-30B	as released	0.352	0.380	0.593	0.363
	$\mathcal{F}_{\text{de}}$	0.356	**0.452**	0.540	**0.457**
	$\mathcal{F}_{\text{mx}}$	0.329	0.424	**0.726**	0.277

Table 6: Population Fidelity Score by question under FA elicitation. The highest score in each column is shown in bold. GPT-5.6 Terra and Muse-Glimmer-30B ($\mathcal{F}_{\text{mx}}$) both score 0.726 on religious attendance after rounding. GPT-3 (archived) denotes the proprietary FA series from the original Machine Bias study, evaluated with the same Population Fidelity metrics on its archived outputs.

Averaged across questions and elicitation modes, only Gemma-4-E4B and Muse-Glimmer-30B improve under both cultural conditions. Gemma-4-E4B has mean ΔPFS of $+0.063$ under $\mathcal{F}_{\text{de}}$ and $+0.050$ under $\mathcal{F}_{\text{mx}}$, while Muse-Glimmer-30B improves by $+0.046$ and $+0.025$, respectively. Qwen3-VL-2B declines by -0.114 and -0.109 and improves in none of its eight $\mathcal{F}_{\text{mx}}$ comparisons. Fine-tuning effects also vary substantially by question: $\mathcal{F}_{\text{de}}$ improves eight of 12 comparisons for happiness but three for social trust, while $\mathcal{F}_{\text{mx}}$ improves two for happiness and seven for political ideology. No model improves on all four questions under both modes for either target.

Those PFS changes hide components that move in different directions. Mean changes over the 48 comparisons, on all demographic cells and for $\mathcal{F}_{\text{de}}$ then $\mathcal{F}_{\text{mx}}$, are -0.077 and $+0.014$ for S_{acc}, -0.103 and -0.064 for S_{adapt}, and $+0.008$ and $+0.004$ for S_{struct}; Section 6.3 gives them within the target countries.

Comparison with archived proprietary-model outputs. The archived outputs from the original Machine Bias study provide a descriptive reference for comparing GPT-5.6 Terra with an earlier proprietary model under FA elicitation. Boelaert et al. (2025) use `davinci-002`, referred to as GPT-3 in the original study, for their FA robustness analysis. We evaluate those archived outputs with the same Population Fidelity metrics used for Terra. This comparison is observational rather than controlled: the models differ in model vintage, post-training, interface, and generation procedure, so observed differences cannot be attributed to any single factor; the corresponding scores appear in Table 6.

Terra attains a higher PFS on three of the four questions: happiness (0.373 vs. 0.245), religious attendance (0.726 vs. 0.000), and social trust (0.416 vs. 0.221). Its PFS is lower for political ideology (0.270 vs. 0.317). The largest difference is in structure, where Terra scores higher on three questions and is nearly unchanged on political ideology.

Figure 7 shows the same comparison within demographic levels. Across 120 level-specific comparisons, Terra has higher structure in 109 and higher PFS in 86, but higher center alignment in only 54 and higher adaptability in 25. Thus, in this descriptive comparison, Terra more often recovers the observed organization of differences among groups, without consistently improving either the amount of between-group variation or agreement with the survey center.

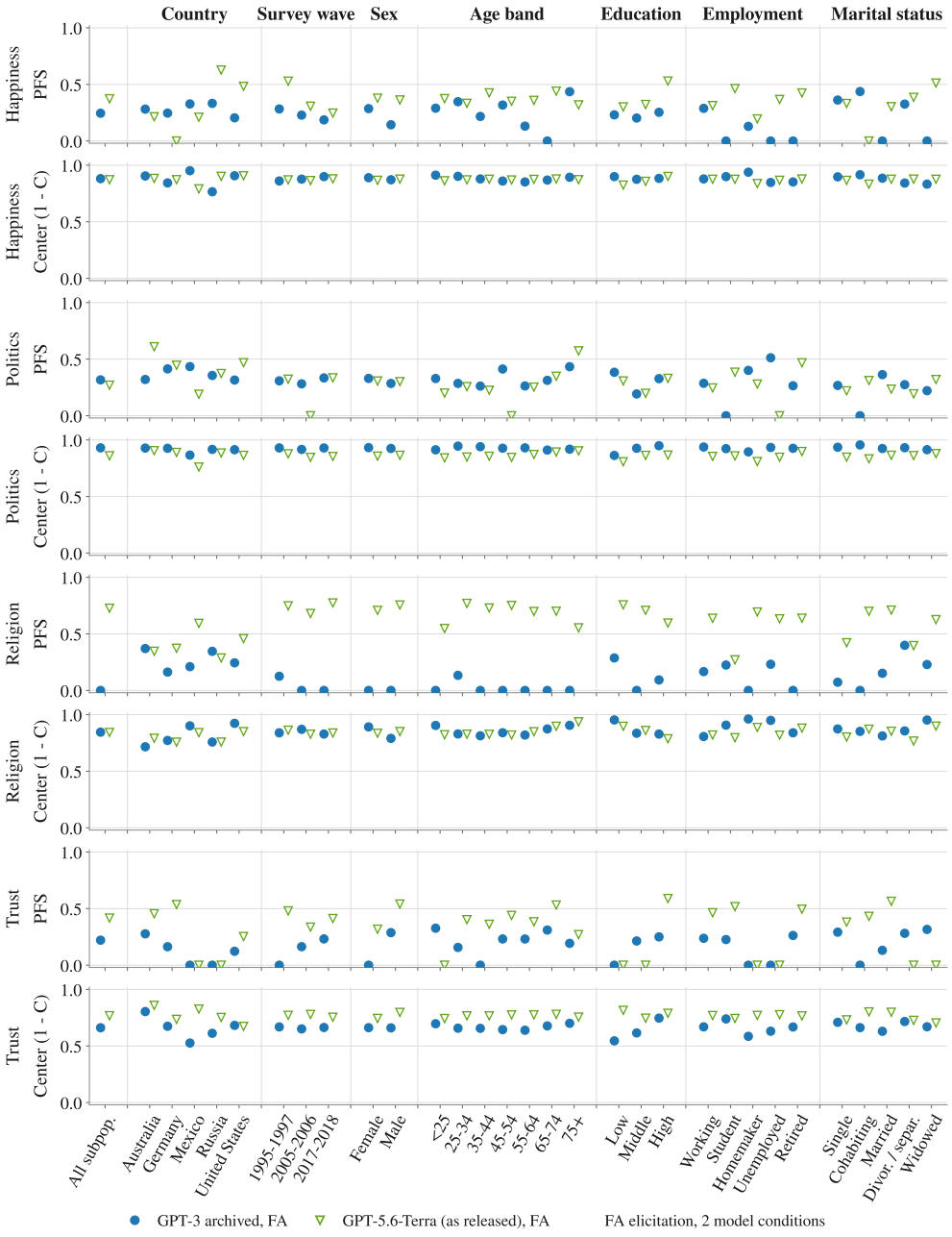


Figure 7: PFS and center alignment by demographic level for the two proprietary models under FA elicitation, shown separately by question. Within each question, the upper row shows PFS and the lower row center alignment; demographic columns follow Figure 3. Blue circles denote archived GPT-3 outputs from the original study, and green triangles denote GPT-5.6 Terra as released.

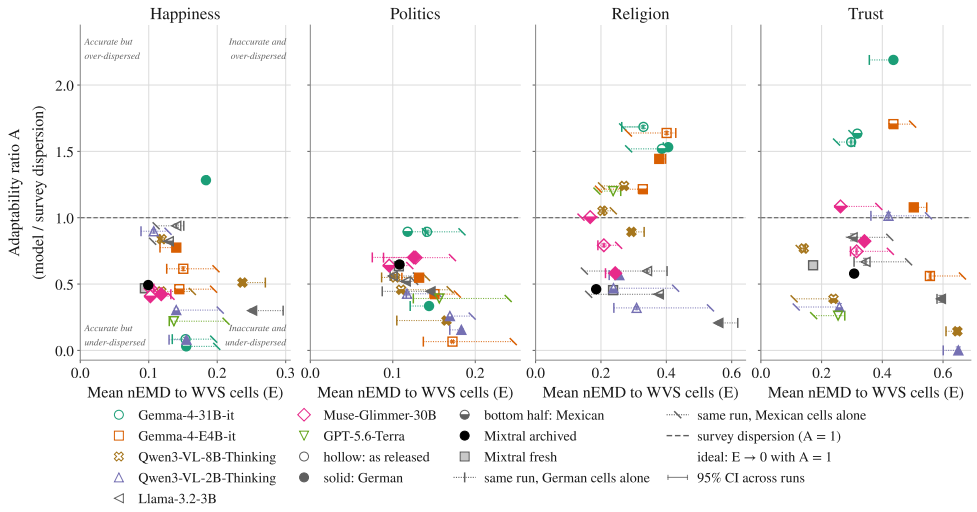


Figure 8: Cell-level error and adaptability under FA elicitation. The horizontal axis reports mean nEMD to the WVS cells, with lower values indicating greater accuracy. The vertical axis reports the adaptability ratio A. Values below $A = 1$ indicate compression of between-group variation, while values above one indicate amplification. Adaptability must be interpreted alongside structure because matching the amount of variation does not establish that it occurs between the corresponding social groups.

D.3 Accuracy and Adaptability by Question

Compression dominates, with 43 of the 52 as-released run–mode combinations at $A < 1$ and only 9 at $A > 1$ (Table 7). All 26 happiness and political ideology combinations compress; amplification occurs only for religious attendance and social trust. Only one of the nine amplifying combinations reaches $\rho > 0.1$, illustrating why adaptability and structure must be examined separately.

Question	Mode	n	$A < 1$	$A > 1$	A range
Happiness	NTP	6	6	0	0.11–0.72
	FA	7	7	0	0.08–0.95
Political ideology	NTP	6	6	0	0.11–0.77
	FA	7	7	0	0.07–0.89
Religious attendance	NTP	6	4	2	0.29–1.61
	FA	7	3	4	0.32–1.68
Social trust	NTP	6	5	1	0.33–1.56
	FA	7	5	2	0.26–1.56
All		52	43	9	0.07–1.68

Table 7: Compression is universal for happiness and political ideology, while amplification occurs only for religious attendance and social trust, as the $A < 1$ and $A > 1$ columns show by counting conditions producing less and more between-group variation than the survey. Counts cover the six open-weight LLMs under both elicitation modes and GPT-5.6 Terra under FA, excluding the Mixtral reference series.

Figure 8 complements these counts by showing how S_{acc} relates to A under FA elicitation. Conditions near $A = 1$ reproduce the survey's overall amount of between-group variation, but this does not establish that the variation is assigned to the corresponding social groups.

D.4 Demographic Breakdown

The demographic-subgroup analysis contains 1,440 paired scores for each cultural target. German fine-tuning improves PFS in 42.5% of these comparisons, reduces it in 49.7%, and leaves 7.8% unchanged. The corresponding proportions for Mexican fine-tuning are 40.8%, 48.5%, and 10.7%.

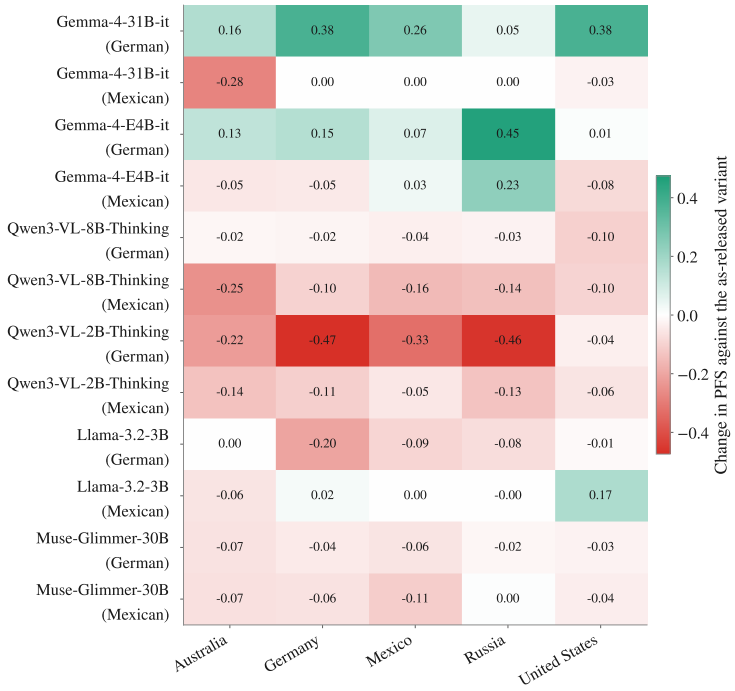


Figure 9: Change in PFS by country after cultural fine-tuning for happiness under FA elicitation. Each row represents a fine-tuned condition and each column a surveyed country. Green indicates improvement relative to the corresponding as-released model, and red indicates decline. For happiness under FA, target-country changes are not consistently more favorable than changes in the other surveyed countries.

Improvements are not concentrated in the target countries. Within Germany, German fine-tuning improves 18 of 48 comparisons, with a mean PFS change of -0.021. Across the other countries, it improves 38.0% of comparisons, with a mean change of -0.019. Within Mexico, Mexican fine-tuning also improves 18 of 48 comparisons, with a mean change of -0.010. Outside Mexico, it improves 32.3%, with a mean change of -0.035.

For each condition and demographic family, we define the within-condition spread as the maximum minus minimum PFS across that family's levels and average this quantity across conditions. The mean within-condition spread in PFS is largest across age (0.287) and employment (0.270), followed by country (0.254), marital status (0.241), education (0.179), survey wave (0.107), and sex (0.054). Although adaptability varies substantially across levels, PFS spread is more sensitive to structure: holding each condition's structure at its family mean reduces the mean spread to between 0.018 and 0.080, whereas holding adaptability at its family mean leaves it between 0.048 and 0.269.

Mean PFS generally increases with age, although not monotonically. It is 0.217 and 0.207 in the two youngest groups and 0.310 and 0.317 in the two oldest. Students have the lowest mean PFS among employment groups (0.171), compared with 0.274 for retired respondents. These estimates should be interpreted cautiously because the oldest and student groups contain comparatively few cells and cell pairs. Table 8 reports the complete results and sample sizes.

Center alignment and PFS by demographic level. Figure 3 compares PFS and center alignment across demographic levels for happiness, while Figure 10 extends the comparison to all four questions. Center alignment varies much less across demographic levels than PFS. Excluding the Mixtral reference series, the median center alignment across the 30 levels ranges from 0.835 for cohabiting respondents to 0.878 for Australia, whereas median PFS ranges from 0.143 for students to 0.343 for respondents aged 75 and older. For happiness alone, the corresponding ranges are 0.848–0.928 for center alignment and 0.178–0.436 for PFS.

Family	Level	PFS	S_{acc}	S_{adapt}	S_{struct}	Cells	Pairs
Country	United States	0.290	0.783	0.508	0.090	140	9,674
	Germany	0.273	0.773	0.527	0.093	144	10,243
	Australia	0.270	0.772	0.549	0.079	136	9,126
	Mexico	0.224	0.749	0.511	0.063	131	8,461
	Russia	0.219	0.755	0.488	0.060	124	7,817
Age	<25	0.217	0.777	0.447	0.079	77	2,941
	25–34	0.207	0.762	0.467	0.072	157	12,238
	35–44	0.235	0.762	0.484	0.101	111	6,149
	45–54	0.264	0.763	0.507	0.111	114	6,505
	55–64	0.263	0.761	0.505	0.093	97	4,655
	65–74	0.310	0.780	0.510	0.117	87	3,713
	75+	0.317	0.786	0.522	0.145	28	373
Sex	Female	0.248	0.771	0.497	0.093	367	67,199
	Male	0.237	0.762	0.498	0.083	306	46,805
Education	Low	0.271	0.760	0.497	0.109	51	1,268
	High	0.215	0.769	0.496	0.076	261	33,965
	Middle	0.205	0.766	0.498	0.066	361	65,196
Employ.	Homemaker	0.275	0.768	0.453	0.120	74	2,718
	Retired	0.274	0.779	0.510	0.093	154	11,819
	Working	0.241	0.758	0.470	0.088	369	68,014
	Unemployed	0.232	0.778	0.476	0.091	33	527
	Student	0.171	0.794	0.436	0.065	38	699
Marital st.	Cohabiting	0.273	0.766	0.375	0.152	38	721
	Married	0.265	0.754	0.448	0.100	360	64,753
	Widowed	0.251	0.787	0.414	0.100	65	2,108
	Div. / sep.	0.233	0.795	0.444	0.079	73	2,631
	Single	0.202	0.776	0.425	0.071	136	9,220
Wave	1995–1997	0.246	0.762	0.507	0.086	221	24,218
	2005–2006	0.250	0.767	0.495	0.100	208	21,839
	2017–2018	0.249	0.771	0.519	0.088	244	29,700

Table 8: Demographic-subgroup Population Fidelity. Metrics are recomputed within each level and then averaged across the 148 run–mode combinations, excluding the archived and fresh Mixtral reference series. Cells and pairs report the mean numbers contributing to each estimate. Estimates based on fewer cells, especially structure, should be interpreted cautiously. Two country-level estimates have zero-variance model distance vectors and receive a structure score of zero under the flatness rule in Appendix B.2.

This difference is also visible within particular demographic levels. Students, unemployed respondents, respondents aged 25–34, and Russian cells have median center alignment within 0.01 of the pooled value (0.860), but substantially lower median PFS (0.143–0.238, compared with 0.254 for the pooled population). Among the 4,440 subgroup-level scores excluding Mixtral references, 295 combine center alignment of at least 0.85 with PFS of at most 0.10. Structure is the limiting component in 278 of these cases. A model can therefore approximate a subgroup center while recovering little of the pattern of differences among the cells that compose it.

Across model conditions, PFS and center alignment are positively but only moderately associated. Excluding Mixtral references, the mean Spearman correlation across demographic levels ranges from 0.263 for political ideology under NTP to 0.425 for social trust under NTP. At the pooled population level, correlations range from 0.302 for religious attendance under FA to 0.674 for political ideology under FA. Bootstrap intervals are wide, particularly at the population level. Pearson correlations give the same qualitative pattern. These results support treating center alignment and Population Fidelity as related but distinct quantities.

The converse pattern also occurs, although less often. A PFS of at least 0.40 together with center alignment of at most 0.75 occurs 84 times, 60 of them for social trust. Fine-tuned conditions account for 60 of these cases. Thus, relatively high subgroup fidelity can also coexist with weaker center alignment. Figure 11 shows that these differences are distributed across model conditions rather than being driven by a single model.

D.5 Additional Population Fidelity Figures

Figure 12 relates the amount of model variation to its correspondence with observed group differences. Figures 14 and 15 extend the main-text fine-tuning and component figures to the remaining

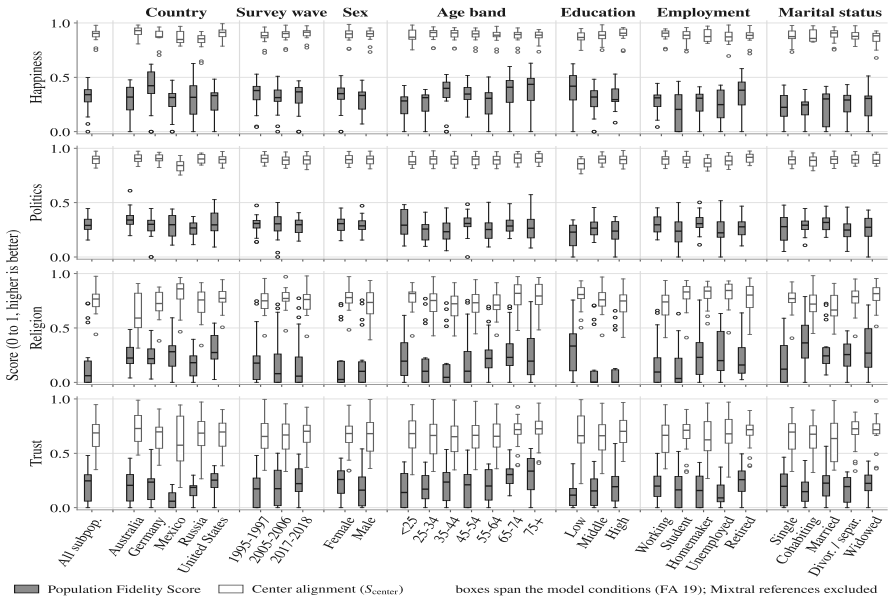


Figure 10: PFS and center alignment by demographic level under FA elicitation, shown separately by question. Filled boxes show the distribution of PFS and hollow boxes the distribution of center alignment across 19 model conditions; Mixtral reference series are excluded. Demographic columns follow Figure 3.

FA questions and the cross-question summary. Complete NTP and FA figures are available in the accompanying repository.

E Robustness and Sensitivity

E.1 Stability Across Repeated Full-Answer Runs

Because FA elicitation is stochastic, repeated runs can produce different sampled responses and therefore slightly different scores. We quantify this run-to-run variation for the 18 locally evaluated conditions—the as-released, $\mathcal{F}_{\text{de}}$, and $\mathcal{F}_{\text{mx}}$ variants of six models—on all four questions, yielding 72 condition–question pairs evaluated three times each. Across repeats, model weights, fine-tuned variants, prompts, decoding settings, and library versions are held fixed; only the random seed stream changes. NTP is deterministic under our decoding configuration and is therefore not repeated.

For each condition–question pair i, we compute the mean score across its $k = 3$ FA runs and a 95% t-interval,

$$\bar{x}_i \pm h_i, \qquad h_i = t_{0.975,\,k-1}\frac{\sigma_i}{\sqrt{k}},$$

where σ_i is the standard deviation across the three runs and h_i is the corresponding CI half-width. Because $k = 3$, $t_{0.975,2} = 4.30$, so individual intervals can be wide even when the observed run-to-run variation is modest.

To summarize run-to-run uncertainty across a set of n condition–question pairs, we pool their variance estimates and report the corresponding CI half-width,

$$h_{\text{pool}} = \frac{t_{0.975,\nu}}{t_{0.975,2}}\sqrt{\frac{1}{n}\sum_{i=1}^{n} h_i^2}, \qquad \nu = n(k-1).$$

Table 9 reports pooled half-widths across the 18 condition–question pairs for each question and across all 72 pairs, while Table 11 reports the individual CI half-widths.

Run-to-run variation is small relative to differences between model conditions. In the variance decomposition, the run-to-run variance component is 2.9% of the between-condition component for

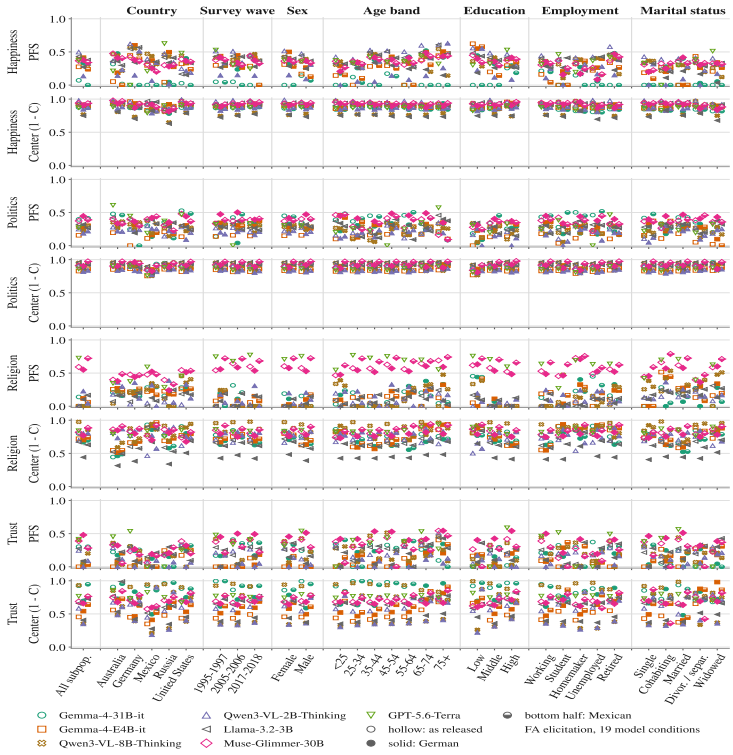


Figure 11: PFS and center alignment by demographic level for individual model conditions under FA elicitation, shown separately by question. Within each question, the upper row shows PFS and the lower row center alignment; demographic columns follow Figure 3. Color and marker shape identify the model. Hollow markers denote as-released conditions, filled markers German fine-tuning, and half-filled markers Mexican fine-tuning.

Question	S_{acc}	S_{adapt}	S_{struct}	PFS	S_{center}	ρ (PFS)
Happiness	0.001	0.013	0.015	0.019	0.001	0.960
Political ideology	0.001	0.008	0.017	0.027	0.001	0.920
Religious attendance	0.002	0.008	0.007	0.029	0.002	0.922
Social trust	0.002	0.021	0.012	0.045	0.002	0.926
All questions	0.002	0.013	0.013	0.031	0.002	0.972

Table 9: Run-to-run uncertainty across repeated FA runs. Entries are pooled 95% CI half-widths, h_{pool}, across the condition–question pairs in each row. The final column reports the mean of the three pairwise Spearman rank correlations in PFS between repeated runs; for "All questions," correlations are computed after pooling the condition–question observations across the four questions. Smaller CI half-widths and larger rank correlations indicate greater stability across repeated elicitation.

PFS, corresponding to an intraclass correlation of 0.971. The corresponding ratios are 1.8% for S_{struct} and below 0.3% for S_{acc}, S_{adapt}, and S_{center}. Thus, accuracy and center alignment are highly stable across repeats, while structure has the largest run-to-run variance relative to between-condition differences among the PFS components.

The qualitative conclusions are similarly stable. In 70 of the 72 condition–question pairs, the same component is limiting in all three runs: S_{struct} in 69 pairs and S_{adapt} in one. This reinforces the main result that structure, rather than accuracy, is typically the limiting dimension of population fidelity. Among the as-released conditions, the adaptability ratio A never crosses one across runs, so no condition switches between compression and amplification. Model-condition rankings are also stable: pairwise Spearman correlations in PFS between runs range from 0.889 to 0.965 across

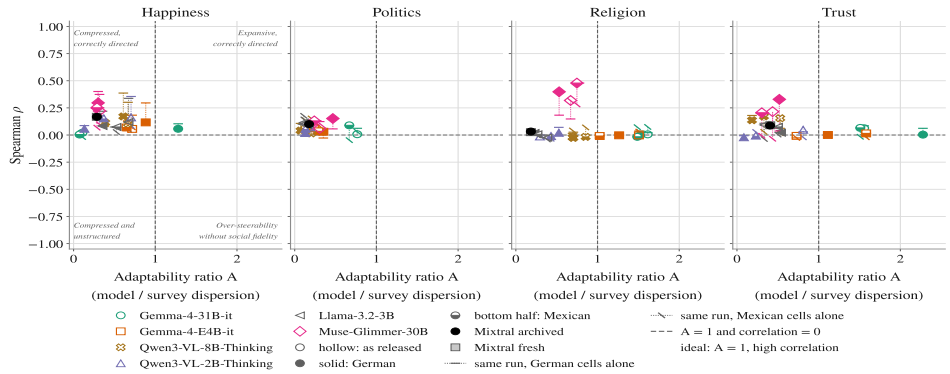
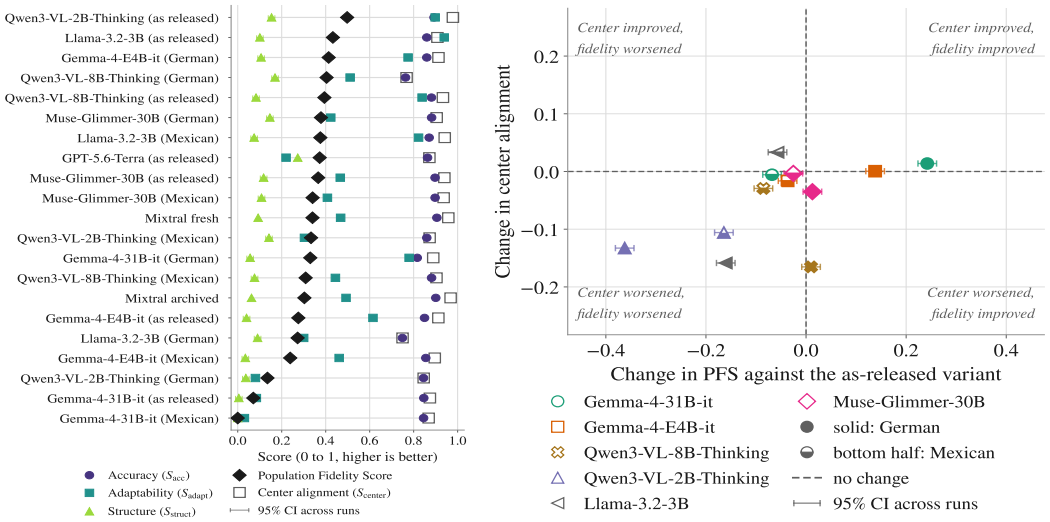


Figure 12: Adaptability and structure under NTP elicitation, shown separately by question. The horizontal axis reports the adaptability ratio A, and the vertical axis the untransformed structure correlation ρ. Dashed lines mark $A = 1$ and $\rho = 0$. Dotted connectors link each population-wide estimate to the corresponding estimate recomputed within the target-country subset: Germany for German fine-tuning and Mexico for Mexican fine-tuning. As-released conditions show both target-country estimates, whereas fine-tuned conditions show only the estimate for their target country. Conditions with $A > 1$ and $\rho \approx 0$ exhibit amplified between-group variation without recovering the observed structure of group differences.



(a) Component profile, all retained cells.

(b) Fine-tuning shifts, all retained cells.

Figure 13: Population Fidelity for happiness under FA elicitation on all retained cells. **(a)** Component scores for all 21 conditions, ordered by PFS; diamonds show PFS and hollow squares center alignment. **(b)** Fine-tuned minus as-released changes computed on the cells retained by both conditions. Figure 2 shows the corresponding target-country results. Plotting conventions are as in Figure 2.

questions. Finally, 63 of the 72 pairs retain identical cell sets in all three runs, so for most pairs the observed movement reflects sampled responses rather than changes in evaluation support.

Much of the remaining PFS sensitivity occurs near the clipping boundary for S_{struct}. Because $S_{\text{struct}} = \max(0, \rho)$, a small run-to-run change that moves ρ across zero can move PFS between zero and a positive value. Nineteen of the 72 condition–question pairs touch this boundary across runs. Near-zero PFS values should therefore not be interpreted as fine-grained rankings: they primarily indicate little or no recovered population structure.

The paired fine-tuning effects are also stable across repeated FA runs (Table 10). Across the 24 model–question comparisons for each cultural target, the run-specific mean ΔPFS on all retained

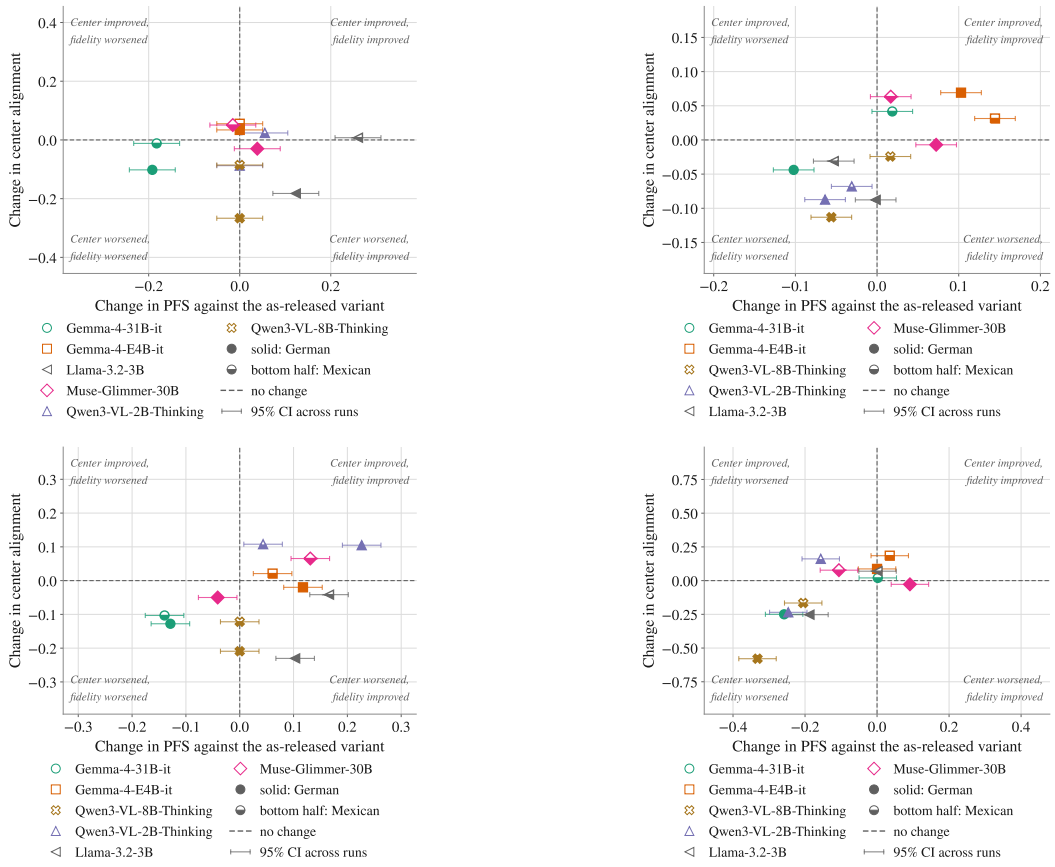


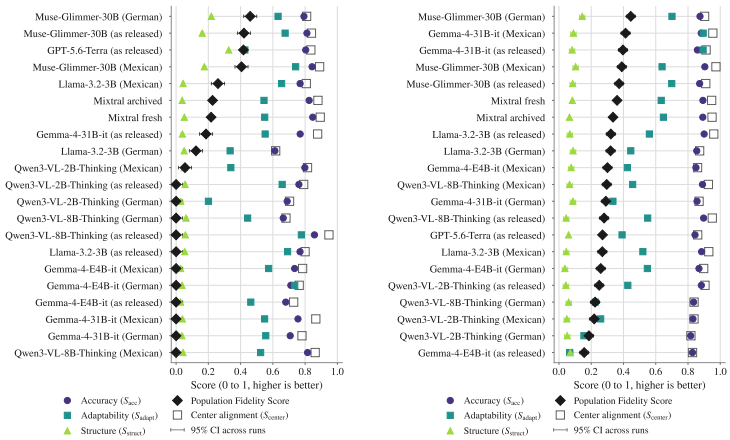
Figure 14: Changes in PFS and center alignment after cultural fine-tuning under FA elicitation, for the cross-question summary, political ideology, religious attendance, and social trust. Each point compares a fine-tuned condition with its corresponding as-released model. Movement to the right indicates improved PFS, and upward movement indicates improved center alignment. The happiness result appears in Figure 2b.

cells and the run-specific mean ΔS_{center} in both scopes vary by at most 0.005. For population-wide ΔPFS, the sign is unchanged across all three runs in 23 of 24 $\mathcal{F}_{\text{de}}$ comparisons and 18 of 24 $\mathcal{F}_{\text{mx}}$ comparisons. The sign of ΔS_{center} is unchanged in all 24 comparisons for each target and scope.

Within the target countries, we summarize individual-comparison run-to-run uncertainty by pooling the 24 model–question-specific variance estimates using the same procedure above ($n = 24$, $\nu = 48$). The resulting pooled 95% CI half-width is 0.073 for ΔPFS in Germany and 0.110 in Mexico. The corresponding pooled half-widths for ΔS_{center} are only 0.005 and 0.003, respectively. Thus, center-alignment shifts are substantially more stable to FA sampling than target-country PFS changes. These intervals quantify uncertainty attributable to repeated FA sampling with the survey, prompts, model parameters, and evaluation definitions held fixed; they are not confidence intervals for the mean effect across model–question comparisons. Appendix E.3 separately examines sensitivity to the metric specification.

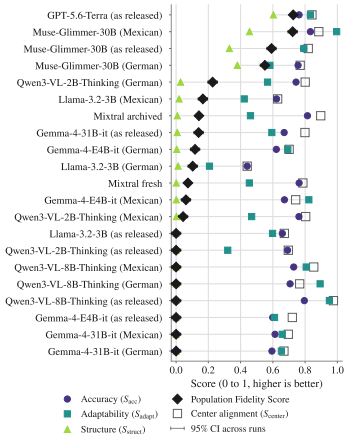
E.2 Agreement Across Elicitation Modes

To assess whether the results depend on how responses are elicited, we compare the rankings of the 20 conditions evaluated under both NTP and FA. Table 12 reports the Spearman correlation between modes for the PFS, its components, and center alignment, calculated separately for each question. Higher correlations indicate that the relative ordering of model conditions is preserved across elicitation methods.

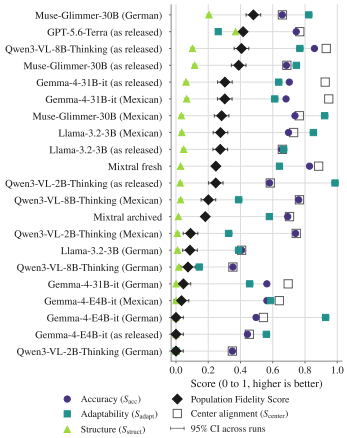


(a) Cross-question summary.

(b) Political ideology.



(c) Religious attendance.



(d) Social trust.

Figure 15: Population Fidelity components under FA elicitation, for the cross-question summary, political ideology, religious attendance, and social trust. Accuracy is generally the strongest component, whereas structure is the weakest and most often limiting; happiness is shown in Figure 2a and movement intervals in Table 9. Fine-tuned conditions are labelled German ($\mathcal{F}_{\text{de}}$) and Mexican ($\mathcal{F}_{\text{mx}}$).

Fine-tuning effects are also directionally similar across elicitation modes. NTP and FA agree on the sign of ΔPFS in 38 of 48 paired comparisons, or 35 of 41 after excluding comparisons with zero change under either mode.

E.3 Metric Sensitivity

The Population Fidelity Score requires several measurement choices, including how between-group dispersion is summarized, how structure is measured, how the three components are combined, and how cells are weighted. We test whether the main conclusions depend on these choices by recomputing the analysis under alternative specifications while holding the underlying model responses, retained cells, and nEMD distances fixed.

Alternative specifications. We consider seven changes individually and one combined specification. For adaptability, we replace the median pairwise distance with the mean, or use the one-sided score $\min(A, 1)$ so that amplification is not penalized. For structure, we replace Spearman's ρ with Pearson's r or Kendall's τ_b. Separately, we test an unclipped version of the default Spearman measure that preserves negative correlations rather than mapping them to zero. We define the

		$\mathcal{F}_{\text{de}}$				$\mathcal{F}_{\text{mx}}$			
		All retained cells		Germany		All retained cells		Mexico	
Model	Question	ΔPFS	ΔS_{center}	ΔPFS	ΔS_{center}	ΔPFS	ΔS_{center}	ΔPFS	ΔS_{center}
Gemma-4-31B-it	Happiness	$+0.242 \pm 0.007$	$+0.014 \pm 0.001$	$+0.376 \pm 0.011$	-0.005 ± 0.002	-0.068 ± 0.025	-0.006 ± 0.001	0.000 ± 0.000	-0.003 ± 0.004
	Political ideology	-0.102 ± 0.057	-0.044 ± 0.001	$+0.278 \pm 0.226$	-0.014 ± 0.001	$+0.019 \pm 0.069$	$+0.042 \pm 0.002$	-0.042 ± 0.385	$+0.087 \pm 0.007$
	Religious attendance	-0.129 ± 0.027	-0.128 ± 0.004	$+0.014 \pm 0.042$	-0.167 ± 0.005	-0.140 ± 0.024	-0.103 ± 0.003	$+0.010 \pm 0.012$	-0.045 ± 0.006
	Social trust	-0.258 ± 0.098	-0.250 ± 0.001	-0.070 ± 0.044	$+0.090 \pm 0.006$	$+0.002 \pm 0.003$	$+0.020 \pm 0.003$	0.000 ± 0.000	-0.065 ± 0.003
	All questions	-0.062 ± 0.010	-0.102 ± 0.001	$+0.150 \pm 0.068$	-0.024 ± 0.003	-0.047 ± 0.014	-0.012 ± 0.001	-0.008 ± 0.095	-0.006 ± 0.003
Gemma-4-E4B-it	Happiness	$+0.138 \pm 0.029$	$+0.001 \pm 0.000$	$+0.154 \pm 0.020$	-0.029 ± 0.003	-0.037 ± 0.014	-0.016 ± 0.000	$+0.032 \pm 0.037$	-0.013 ± 0.003
	Political ideology	$+0.103 \pm 0.066$	$+0.069 \pm 0.003$	$+0.307 \pm 0.138$	$+0.041 \pm 0.004$	$+0.144 \pm 0.023$	$+0.031 \pm 0.001$	$+0.086 \pm 0.437$	$+0.082 \pm 0.001$
	Religious attendance	$+0.118 \pm 0.030$	-0.020 ± 0.003	-0.026 ± 0.008	-0.017 ± 0.005	$+0.061 \pm 0.146$	$+0.021 \pm 0.005$	-0.051 ± 0.121	$+0.079 \pm 0.003$
	Social trust	0.000 ± 0.000	$+0.087 \pm 0.004$	$+0.021 \pm 0.090$	$+0.014 \pm 0.002$	$+0.035 \pm 0.152$	$+0.185 \pm 0.006$	0.000 ± 0.000	$+0.172 \pm 0.010$
	All questions	$+0.090 \pm 0.014$	$+0.034 \pm 0.001$	$+0.114 \pm 0.062$	$+0.002 \pm 0.002$	$+0.051 \pm 0.026$	$+0.055 \pm 0.002$	$+0.017 \pm 0.095$	$+0.080 \pm 0.003$
Qwen3-VL-8B-Thinking	Happiness	$+0.010 \pm 0.006$	-0.165 ± 0.002	-0.023 ± 0.059	-0.165 ± 0.007	-0.085 ± 0.079	-0.029 ± 0.004	-0.158 ± 0.106	-0.069 ± 0.003
	Political ideology	-0.056 ± 0.025	-0.113 ± 0.003	-0.080 ± 0.106	-0.038 ± 0.004	$+0.016 \pm 0.072$	-0.024 ± 0.002	-0.044 ± 0.109	-0.042 ± 0.002
	Religious attendance	0.000 ± 0.000	-0.209 ± 0.007	$+0.010 \pm 0.089$	-0.203 ± 0.013	0.000 ± 0.000	-0.122 ± 0.008	$+0.048 \pm 0.125$	$+0.079 \pm 0.014$
	Social trust	-0.332 ± 0.193	-0.579 ± 0.011	-0.288 ± 0.344	-0.517 ± 0.022	-0.205 ± 0.055	-0.166 ± 0.005	$+0.049 \pm 0.311$	-0.039 ± 0.004
	All questions	-0.094 ± 0.053	-0.267 ± 0.002	-0.095 ± 0.089	-0.231 ± 0.005	-0.069 ± 0.027	-0.085 ± 0.001	-0.026 ± 0.127	-0.018 ± 0.004
Qwen3-VL-2B-Thinking	Happiness	-0.362 ± 0.035	-0.133 ± 0.002	-0.474 ± 0.077	-0.087 ± 0.011	-0.164 ± 0.032	-0.106 ± 0.001	-0.054 ± 0.217	-0.104 ± 0.010
	Political ideology	-0.064 ± 0.060	-0.087 ± 0.005	-0.027 ± 0.166	-0.063 ± 0.004	-0.031 ± 0.062	-0.068 ± 0.005	-0.047 ± 0.215	-0.052 ± 0.005
	Religious attendance	$+0.227 \pm 0.003$	$+0.105 \pm 0.005$	$+0.324 \pm 0.106$	$+0.061 \pm 0.015$	$+0.043 \pm 0.187$	$+0.108 \pm 0.002$	-0.046 ± 0.200	$+0.107 \pm 0.007$
	Social trust	-0.247 ± 0.048	-0.235 ± 0.009	-0.266 ± 0.171	-0.246 ± 0.027	-0.156 ± 0.185	$+0.161 \pm 0.014$	$+0.034 \pm 0.575$	$+0.453 \pm 0.007$
	All questions	-0.111 ± 0.011	-0.088 ± 0.002	-0.111 ± 0.032	-0.084 ± 0.009	-0.077 ± 0.074	$+0.024 \pm 0.003$	-0.028 ± 0.130	$+0.101 \pm 0.001$
Llama-3.2-3B	Happiness	-0.160 ± 0.052	-0.159 ± 0.005	-0.204 ± 0.044	-0.165 ± 0.009	-0.057 ± 0.064	$+0.033 \pm 0.002$	$+0.003 \pm 0.053$	-0.011 ± 0.006
	Political ideology	-0.002 ± 0.014	-0.088 ± 0.003	-0.198 ± 0.346	$+0.006 \pm 0.021$	-0.053 ± 0.048	-0.031 ± 0.003	-0.056 ± 0.146	$+0.012 \pm 0.014$
	Religious attendance	$+0.103 \pm 0.030$	-0.230 ± 0.006	$+0.077 \pm 0.323$	-0.218 ± 0.009	$+0.166 \pm 0.081$	-0.042 ± 0.004	$+0.080 \pm 0.343$	-0.028 ± 0.005
	Social trust	-0.188 ± 0.164	-0.253 ± 0.001	-0.211 ± 0.133	-0.280 ± 0.014	$+0.001 \pm 0.044$	$+0.070 \pm 0.009$	$+0.062 \pm 0.393$	$+0.063 \pm 0.011$
	All questions	-0.062 ± 0.042	-0.182 ± 0.001	-0.134 ± 0.069	-0.164 ± 0.005	$+0.014 \pm 0.027$	$+0.008 \pm 0.002$	$+0.022 \pm 0.053$	$+0.009 \pm 0.004$
Muse-Glimmer-30B	Happiness	$+0.013 \pm 0.030$	-0.035 ± 0.001	-0.045 ± 0.096	-0.030 ± 0.004	-0.025 ± 0.010	-0.003 ± 0.003	-0.112 ± 0.136	-0.044 ± 0.002
	Political ideology	$+0.072 \pm 0.023$	-0.007 ± 0.005	-0.040 ± 0.034	$+0.022 \pm 0.010$	$+0.017 \pm 0.057$	$+0.063 \pm 0.007$	$+0.042 \pm 0.191$	$+0.072 \pm 0.006$
	Religious attendance	-0.041 ± 0.031	-0.050 ± 0.005	-0.037 ± 0.074	-0.038 ± 0.001	$+0.131 \pm 0.007$	$+0.065 \pm 0.002$	$+0.004 \pm 0.037$	$+0.160 \pm 0.003$
	Social trust	$+0.091 \pm 0.030$	-0.028 ± 0.000	$+0.031 \pm 0.183$	-0.048 ± 0.004	-0.106 ± 0.043	$+0.077 \pm 0.004$	$+0.104 \pm 0.125$	$+0.043 \pm 0.007$
	All questions	$+0.034 \pm 0.019$	-0.030 ± 0.002	-0.023 ± 0.036	-0.024 ± 0.003	$+0.004 \pm 0.027$	$+0.051 \pm 0.001$	$+0.010 \pm 0.005$	$+0.058 \pm 0.001$
All models	Happiness	-0.020 ± 0.013	-0.080 ± 0.001	-0.036 ± 0.003	-0.080 ± 0.001	-0.073 ± 0.011	-0.021 ± 0.001	-0.048 ± 0.023	-0.041 ± 0.003
	Political ideology	-0.008 ± 0.005	-0.045 ± 0.002	$+0.040 \pm 0.039$	-0.007 ± 0.003	$+0.019 \pm 0.007$	$+0.002 \pm 0.000$	-0.010 ± 0.091	$+0.026 \pm 0.004$
	Religious attendance	$+0.046 \pm 0.007$	-0.089 ± 0.002	$+0.060 \pm 0.065$	-0.097 ± 0.001	$+0.044 \pm 0.035$	-0.012 ± 0.002	$+0.008 \pm 0.082$	$+0.059 \pm 0.003$
	Social trust	-0.156 ± 0.027	-0.210 ± 0.001	-0.130 ± 0.061	-0.165 ± 0.002	-0.071 ± 0.042	$+0.058 \pm 0.005$	$+0.042 \pm 0.073$	$+0.104 \pm 0.006$
	All questions	-0.034 ± 0.006	-0.106 ± 0.001	-0.017 ± 0.020	-0.087 ± 0.000	-0.020 ± 0.005	$+0.007 \pm 0.001$	-0.002 ± 0.020	$+0.037 \pm 0.002$

Table 10: Paired fine-tuning effects across the three repeated FA runs. Each model–question entry reports the mean fine-tuned-minus-as-released change across the three runs $\pm$ its 95% confidence interval, computed on the cells retained in all runs of both conditions. "All questions" averages the four question-specific changes within each run before constructing the interval, while "All models" averages the six model-specific changes within each run. Positive values indicate increases relative to the corresponding as-released condition.

corresponding signed score as

$$\text{PFS}_{\text{signed}} = \text{sign}(\rho)\,(S_{\text{acc}} \cdot S_{\text{adapt}} \cdot |\rho|)^{1/3},$$

where $\text{sign}(\rho)$ is -1, 0, or 1 according to whether ρ is negative, zero, or positive. Thus, the magnitude of the score is computed from $|\rho|$, while its sign preserves the direction of the structure correlation. Unlike the default PFS, this sensitivity variant can take negative values.

We also replace the geometric mean with the arithmetic mean and, separately, weight accuracy and center alignment by the number of survey respondents in each cell. The combined specification uses mean pairwise distance, Pearson structure without clipping, and respondent weighting. Table 13 compares these alternatives with the default definition.

Main conclusions are stable. The central results are largely unchanged across specifications. Accuracy remains substantially higher than structure: mean accuracy ranges only from 0.773 to 0.775, whereas mean structure ranges from 0.055 to 0.098. Structure remains the limiting component in at least 156 of 164 population-level evaluations and 4,486 of 4,920 subgroup-level evaluations; accuracy is never limiting. Among specifications retaining the geometric mean, PFS rankings remain highly similar to the default ($\rho \geq 0.979$).

At the population-wide scope, neither CultureLLM-style condition improves PFS on average under any specification. Among geometric-mean specifications, mean paired changes range from -0.028 to -0.016 for German fine-tuning and from -0.033 to -0.012 for Mexican fine-tuning. Target-country results are somewhat more specification-sensitive. Within Mexico, center alignment improves by approximately 0.038 across specifications, whereas target-country PFS is negative under all but the unclipped specification, where the mean change is effectively zero. Within Germany, target-country PFS is negative under the default and most alternative specifications but becomes positive under the combined specification ($\Delta\text{PFS} = +0.016$). Thus, the population-wide conclusion is robust, while the sign of a target-country mean can depend on the metric specification.

		As released					$\mathcal{F}_{\text{de}}$					$\mathcal{F}_{\text{mx}}$				
Model	Question	Acc.	Adapt.	Struct.	PFS	Center	Acc.	Adapt.	Struct.	PFS	Center	Acc.	Adapt.	Struct.	PFS	Center
Gemma-4-31B-it	Happiness	0.000	0.018	0.005	0.023	0.001	0.000	0.025	0.003	0.002	0.002	0.001	0.065	0.000	0.000	0.001
	Political ideology	0.001	0.006	0.020	0.034	0.001	0.001	0.023	0.034	0.044	0.002	0.002	0.020	0.034	0.053	0.003
	Religious attendance	0.003	0.002	0.003	0.025	0.003	0.001	0.013	0.000	0.000	0.002	0.002	0.007	0.000	0.000	0.001
	Social trust	0.003	0.030	0.005	0.009	0.001	0.005	0.016	0.002	0.102	0.007	0.001	0.031	0.005	0.005	0.004
	All questions	0.001	0.009	0.006	0.013	0.001	0.001	0.011	0.009	0.030	0.002	0.001	0.020	0.009	0.014	0.001
Gemma-4-E4B-it	Happiness	0.001	0.014	0.008	0.020	0.001	0.002	0.019	0.007	0.010	0.001	0.000	0.016	0.013	0.031	0.001
	Political ideology	0.000	0.002	0.028	0.021	0.000	0.003	0.022	0.019	0.046	0.003	0.001	0.013	0.022	0.031	0.001
	Religious attendance	0.001	0.019	0.000	0.000	0.002	0.002	0.007	0.003	0.030	0.004	0.003	0.012	0.004	0.146	0.005
	Social trust	0.007	0.089	0.000	0.000	0.006	0.003	0.032	0.000	0.000	0.003	0.004	0.022	0.005	0.147	0.003
	All questions	0.002	0.025	0.008	0.008	0.002	0.001	0.012	0.005	0.015	0.002	0.001	0.009	0.007	0.057	0.002
Qwen3-VL-8B-Thinking	Happiness	0.001	0.019	0.030	0.051	0.002	0.003	0.023	0.049	0.046	0.004	0.001	0.021	0.054	0.074	0.002
	Political ideology	0.002	0.029	0.010	0.022	0.003	0.001	0.005	0.028	0.036	0.001	0.002	0.016	0.063	0.093	0.001
	Religious attendance	0.006	0.027	0.000	0.000	0.004	0.003	0.034	0.000	0.000	0.003	0.004	0.018	0.000	0.000	0.008
	Social trust	0.008	0.078	0.040	0.052	0.008	0.003	0.007	0.040	0.160	0.003	0.005	0.025	0.033	0.073	0.004
	All questions	0.003	0.024	0.014	0.020	0.003	0.001	0.011	0.018	0.046	0.001	0.002	0.011	0.024	0.037	0.003
Qwen3-VL-2B-Thinking	Happiness	0.003	0.043	0.031	0.033	0.003	0.001	0.016	0.024	0.036	0.001	0.001	0.027	0.029	0.033	0.002
	Political ideology	0.001	0.021	0.044	0.081	0.004	0.000	0.011	0.022	0.022	0.000	0.001	0.007	0.012	0.020	0.000
	Religious attendance	0.005	0.001	0.000	0.000	0.004	0.002	0.009	0.001	0.003	0.002	0.004	0.020	0.009	0.187	0.002
	Social trust	0.012	0.050	0.015	0.048	0.011	0.002	0.000	0.000	0.000	0.002	0.005	0.031	0.034	0.230	0.004
	All questions	0.004	0.018	0.015	0.027	0.003	0.001	0.006	0.009	0.011	0.001	0.002	0.012	0.013	0.080	0.001
Llama-3.2-3B	Happiness	0.003	0.041	0.034	0.052	0.005	0.001	0.011	0.026	0.023	0.000	0.002	0.018	0.031	0.052	0.004
	Political ideology	0.002	0.009	0.047	0.080	0.000	0.004	0.023	0.053	0.072	0.003	0.001	0.018	0.054	0.126	0.003
	Religious attendance	0.009	0.026	0.000	0.000	0.007	0.002	0.015	0.010	0.030	0.002	0.004	0.012	0.023	0.081	0.004
	Social trust	0.003	0.049	0.014	0.028	0.004	0.002	0.027	0.023	0.189	0.002	0.005	0.054	0.024	0.054	0.006
	All questions	0.003	0.019	0.016	0.027	0.002	0.001	0.011	0.017	0.055	0.001	0.002	0.016	0.019	0.045	0.002
Muse-Glimmer-30B	Happiness	0.002	0.036	0.047	0.047	0.002	0.001	0.007	0.056	0.052	0.001	0.002	0.010	0.034	0.038	0.001
	Political ideology	0.007	0.009	0.010	0.016	0.006	0.002	0.016	0.023	0.027	0.004	0.002	0.014	0.058	0.074	0.004
	Religious attendance	0.002	0.010	0.015	0.006	0.003	0.007	0.018	0.049	0.030	0.007	0.001	0.005	0.014	0.008	0.001
	Social trust	0.002	0.056	0.047	0.061	0.002	0.003	0.036	0.055	0.050	0.002	0.006	0.046	0.013	0.030	0.003
	All questions	0.002	0.018	0.018	0.021	0.002	0.002	0.012	0.026	0.022	0.002	0.002	0.013	0.019	0.024	0.001
All models	Happiness	0.001	0.016	0.015	0.020	0.001	0.001	0.009	0.017	0.017	0.001	0.001	0.016	0.016	0.022	0.001
	Political ideology	0.002	0.008	0.015	0.025	0.002	0.001	0.009	0.016	0.022	0.001	0.001	0.008	0.023	0.038	0.001
	Religious attendance	0.003	0.009	0.003	0.005	0.002	0.002	0.009	0.010	0.011	0.002	0.002	0.007	0.006	0.052	0.002
	Social trust	0.003	0.031	0.013	0.020	0.003	0.002	0.012	0.015	0.056	0.002	0.002	0.019	0.011	0.060	0.002
	All questions	0.002	0.017	0.012	0.018	0.002	0.001	0.009	0.014	0.029	0.001	0.001	0.012	0.014	0.042	0.002

Table 11: Run-to-run uncertainty by model, question, and condition. Question-specific entries are 95% CI half-widths for the mean score across the three FA runs, $t_{0.975,\,k-1}\sigma_i/\sqrt{k}$. "All questions" and "All models" entries pool the corresponding run-level variance estimates using their combined degrees of freedom.

Question	PFS	S_{struct}	S_{acc}	S_{adapt}	S_{center}
Happiness	0.838	0.791	0.968	0.788	0.925
Political ideology	0.695	0.549	0.744	0.567	0.792
Religious attendance	0.880	0.926	0.968	0.913	0.925
Social trust	0.864	0.823	0.974	0.502	0.986

Table 12: Agreement between NTP and FA elicitation. Each entry is the Spearman rank correlation between the two modes across the 20 conditions evaluated under both. The structure column uses the untransformed correlation ρ.

Where the specification matters. Some choices affect the numerical scale without changing the main conclusions. Replacing median with mean pairwise distance raises average adaptability from 0.496 to 0.521. Pearson correlation produces somewhat higher structure scores than Spearman, while Kendall correlation produces somewhat lower scores. Removing the zero clipping converts most nonpositive-structure cases into negative PFS values rather than zeros.

The largest change comes from replacing the geometric mean with an arithmetic mean. Mean PFS rises from 0.243 to 0.451, all exact zeros disappear, and the ranking correlation with the default falls to 0.669. This occurs because the arithmetic mean allows high accuracy and adaptability to compensate for very low structure, whereas the geometric mean treats the three dimensions as jointly required.

Finally, paired fine-tuning comparisons are evaluated on the intersection of cells retained by the fine-tuned and as-released conditions. Scoring each condition on its own retained cells produces slightly different changes because response coverage itself then affects adaptability and structure. Across the sensitivity specifications, this alternative does not change the sign of the pooled mean fine-tuning effects. We therefore retain the common-cell construction so that paired differences reflect changes in model responses rather than changes in evaluation support.

	Mean / median				Stability				Fine-tuning ΔPFS		Target-country ΔPFS / ΔS_{center}	
Variant	Acc.	Adapt.	Struct.	PFS	Struct. limit	Struct. limit (4,920)	PFS ≤ 0	ρ vs. default	German	Mexican	Germany	Mexico
Default (Section 4)	0.775 / 0.832	0.496 / 0.497	0.082 / 0.060	0.243 / 0.254	159	4,554	30	1.000	−0.024 (21/23)	−0.014 (19/23)	−0.021 / −0.084	−0.010 / +0.038
Mean pairwise distance	0.775 / 0.832	0.521 / 0.508	0.082 / 0.060	0.248 / 0.265	164	4,679	30	0.995	−0.024 (22/22)	−0.017 (18/24)	−0.017 / −0.084	−0.009 / +0.038
One-sided $\min(A, 1)$	0.775 / 0.832	0.545 / 0.519	0.082 / 0.060	0.246 / 0.254	159	4,558	30	0.997	−0.025 (21/23)	−0.014 (19/23)	−0.020 / −0.084	−0.011 / +0.038
Pearson r	0.775 / 0.832	0.496 / 0.497	0.098 / 0.067	0.257 / 0.262	156	4,486	27	0.988	−0.028 (20/24)	−0.025 (16/26)	−0.018 / −0.084	−0.016 / +0.038
Kendall τ_b	0.775 / 0.832	0.496 / 0.497	0.055 / 0.040	0.213 / 0.222	162	4,714	30	1.000	−0.021 (21/23)	−0.012 (19/23)	−0.018 / −0.084	−0.009 / +0.038
Unclipped structure	0.775 / 0.832	0.496 / 0.497	0.080 / 0.060	0.216 / 0.254	160	4,596	30	0.997	−0.016 (22/26)	−0.016 (22/26)	−0.012 / −0.084	−0.001 / +0.038
Arithmetic mean	0.775 / 0.832	0.496 / 0.497	0.082 / 0.060	0.451 / 0.455	159	4,554	0	0.669	−0.058 (18/30)	−0.015 (21/27)	−0.071 / −0.084	−0.020 / +0.038
Respondent-weighted accuracy and center	0.773 / 0.833	0.496 / 0.497	0.082 / 0.060	0.243 / 0.255	159	4,554	30	1.000	−0.024 (21/23)	−0.014 (19/23)	−0.021 / −0.090	−0.010 / +0.038
Combined	0.773 / 0.833	0.521 / 0.508	0.094 / 0.067	0.232 / 0.263	159	4,606	27	0.979	−0.023 (22/26)	−0.033 (16/32)	+0.016 / −0.090	−0.017 / +0.038

Table 13: Sensitivity of Population Fidelity to alternative metric specifications. Acc., Adapt., Struct., and PFS report the mean and median across the 164 evaluated combinations. "Struct. limit" reports how often structure is the lowest PFS component, first at the population level and then across 4,920 subgroup-level scores, including the Mixtral reference series. ρ vs. default is the Spearman correlation between PFS under each specification and under the default. Fine-tuning columns report mean paired changes in PFS; parentheses give the numbers of increases and decreases. The final columns report mean changes in PFS and center alignment within the corresponding target country. For specifications without clipping, the structure value retains its sign and PFS may be negative.

F Distribution-Matched and Subgroup-Matched Fine-Tuning

The CultureLLM-style conditions used in the main analysis train toward a single target answer per question. The reduction in between-group variation reported in Section 6 may therefore reflect that objective rather than survey-based adaptation in general. To examine whether Population Fidelity distinguishes alternative adaptation objectives, we evaluate two additional methods on a common backbone. We compare the as-released model with four adaptation conditions: CultureLLM-style German fine-tuning ($\mathcal{F}_{\text{de}}$), CultureLLM-style Mexican fine-tuning ($\mathcal{F}_{\text{mx}}$), distribution matching (SimLLCultureDist), and subgroup matching (SubPOP).

SimLLCultureDist (Cao et al., 2025) trains toward aggregate country-level response distributions. These targets contain country-level response distributions but no within-country demographic subgroup information. SubPOP (Suh et al., 2025), by contrast, trains toward response distributions for demographic subgroups and therefore explicitly incorporates subgroup-level information. These conditions provide complementary examples of how fine-tuning approaches with different supervision targets score on the PFS components.

We fine-tune Qwen3-VL-2B-Thinking using the released data and training recipes of SimLLCultureDist and SubPOP. SimLLCultureDist contains 6,841 country-level response distributions from 46 countries and uses a first-token KL-divergence objective. SubPOP-Train contains 71,026 subgroup-level response distributions covering 3,229 American Trends Panel questions and 22 United States subgroups, using the same loss. We evaluate both conditions using the same procedure as the other model conditions. Table 14 reports first-run scores; neither condition is included among the 164 combinations analyzed in the main text.

The training data also differ in their overlap with the evaluation items. The distribution-matched data contain the happiness and social-trust items for four of the five evaluation countries. SubPOP-Train contains eight religious-attendance items and one social-trust item for United States subgroups only. Happiness and political ideology are absent from SubPOP-Train, and neither Germany nor Mexico is represented as a subgroup-training target.

The distribution-matched condition attains the highest median center alignment of the five conditions on this backbone (0.883 versus 0.827–0.843) and improves on the as-released model in seven of eight question–mode combinations. PFS gives a more mixed result: it improves in three combinations, declines in three, and remains zero in two, with median PFS increasing from 0.259 to 0.330.

Center alignment can also obscure large differences in the PFS components. For social trust under NTP, $\mathcal{F}_{\text{mx}}$ and the distribution-matched condition attain nearly identical center alignment (0.815 and 0.814), but their PFS values are 0.000 and 0.425, respectively. $\mathcal{F}_{\text{mx}}$ strongly compresses between-group variation ($A = 0.231$) and recovers no positive structure ($S_{\text{struct}} = 0$), whereas the distribution-matched condition preserves substantially more variation ($A = 0.818$) and positive structure ($S_{\text{struct}} = 0.119$). Thus, similar center alignment can coexist with very different representations of variation among groups (Figure 16).

		FA					NTP				
Question	Condition	S_{acc}	A	S_{struct}	PFS	S_{center}	S_{acc}	A	S_{struct}	PFS	S_{center}
Happiness	As released	0.892	0.913	0.143	0.488	0.979	0.894	0.704	0.162	0.467	0.962
	German ($\mathcal{F}_{\mathrm{de}}$)	0.845	0.077	0.044	0.141	0.846	0.856	0.133	0.058	0.187	0.859
	Mexican ($\mathcal{F}_{\mathrm{mx}}$)	0.860	0.308	0.146	0.338	0.873	0.879	0.371	0.154	0.369	0.905
	Distribution-matched	0.894	0.723	0.129	0.437	0.953	0.899	0.598	0.140	0.422	0.965
	Subgroup-matched	0.898	0.485	0.036	0.250	0.952	0.897	0.192	0.113	0.269	0.971
Political ideology	As released	0.883	0.419	0.063	0.285	0.905	0.904	0.189	0.070	0.229	0.948
	German ($\mathcal{F}_{\mathrm{de}}$)	0.817	0.150	0.060	0.194	0.818	0.859	0.129	0.018	0.127	0.867
	Mexican ($\mathcal{F}_{\mathrm{mx}}$)	0.831	0.262	0.053	0.226	0.837	0.848	0.132	0.057	0.185	0.855
	Distribution-matched	0.883	0.501	0.049	0.278	0.910	0.902	0.219	0.074	0.245	0.959
	Subgroup-matched	0.835	0.525	0.047	0.274	0.858	0.872	0.121	0.028	0.144	0.906
Religious attendance	As released	0.690	0.320	0.000	0.000	0.695	0.762	0.295	0.000	0.000	0.779
	German ($\mathcal{F}_{\mathrm{de}}$)	0.744	0.567	0.027	0.225	0.801	0.777	0.526	0.021	0.204	0.837
	Mexican ($\mathcal{F}_{\mathrm{mx}}$)	0.761	0.475	0.000	0.000	0.803	0.796	0.431	0.000	0.000	0.844
	Distribution-matched	0.763	0.487	0.000	0.000	0.784	0.813	0.319	0.000	0.000	0.856
	Subgroup-matched	0.751	0.280	0.005	0.102	0.764	0.796	0.138	0.005	0.082	0.828
Social trust	As released	0.585	1.038	0.022	0.233	0.589	0.632	0.810	0.048	0.291	0.636
	German ($\mathcal{F}_{\mathrm{de}}$)	0.350	0.000	0.000	0.000	0.350	0.359	0.085	0.000	0.000	0.359
	Mexican ($\mathcal{F}_{\mathrm{mx}}$)	0.741	0.316	0.000	0.000	0.745	0.802	0.231	0.000	0.000	0.815
	Distribution-matched	0.776	1.251	0.091	0.383	0.819	0.791	0.818	0.119	0.425	0.814
	Subgroup-matched	0.537	0.509	0.005	0.111	0.537	0.605	0.246	0.046	0.190	0.605
Mean change from as released	German ($\mathcal{F}_{\mathrm{de}}$)	-0.073	-0.474	-0.024	-0.111	-0.088	-0.085	-0.281	-0.046	-0.118	-0.101
	Mexican ($\mathcal{F}_{\mathrm{mx}}$)	$+0.036$	-0.332	-0.007	-0.110	$+0.023$	$+0.033$	-0.208	-0.018	-0.108	$+0.023$
	Distribution-matched	$+0.066$	$+0.068$	$+0.010$	$+0.023$	$+0.074$	$+0.053$	-0.011	$+0.013$	$+0.026$	$+0.067$
	Subgroup-matched	-0.007	-0.223	-0.034	-0.067	-0.015	-0.006	-0.325	-0.022	-0.076	-0.004

Both modes pooled		S_{acc}	A	S_{struct}	PFS	S_{center}
Median over the eight combinations	As released	0.823	0.561	0.055	0.259	0.842
	German ($\mathcal{F}_{\mathrm{de}}$)	0.797	0.131	0.024	0.164	0.827
	Mexican ($\mathcal{F}_{\mathrm{mx}}$)	0.816	0.312	0.027	0.093	0.841
	Distribution-matched	0.848	0.549	0.082	0.330	0.883
	Subgroup-matched	0.815	0.263	0.032	0.167	0.843

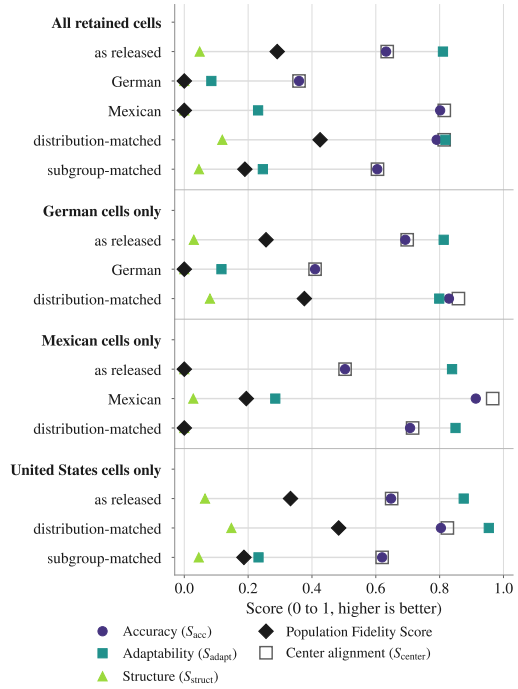
Table 14: Population Fidelity diagnostics for Qwen3-VL-2B-Thinking under CultureLLM-style German fine-tuning ($\mathcal{F}_{\mathrm{de}}$), CultureLLM-style Mexican fine-tuning ($\mathcal{F}_{\mathrm{mx}}$), distribution matching (SimLLCultureDist), and subgroup matching (SubPOP). Scores are from the first run on all retained cells (687 cells; 639 for political ideology). A is the raw adaptability ratio; PFS uses the corresponding symmetric adaptability score S_{adapt}. "Mean change from as released" averages the within-question difference from the as-released condition across the four questions. The final five rows report the median of each condition across its eight question–mode combinations.

The subgroup-matched condition does not consistently recover group structure in this evaluation. Its median adaptability ratio is $A = 0.263$, median $S_{\mathrm{struct}} = 0.032$, and structure is the limiting component in all eight question–mode combinations. PFS declines relative to the as-released model in six of eight combinations. Its two improvements occur for religious attendance, which has the greatest item-level overlap with the SubPOP training data among our evaluation questions. For happiness under FA, for example, the distribution-matched and subgroup-matched conditions have nearly identical center alignment (0.953 and 0.952) but PFS values of 0.437 and 0.250, respectively, reflecting lower adaptability and structure under subgroup matching.

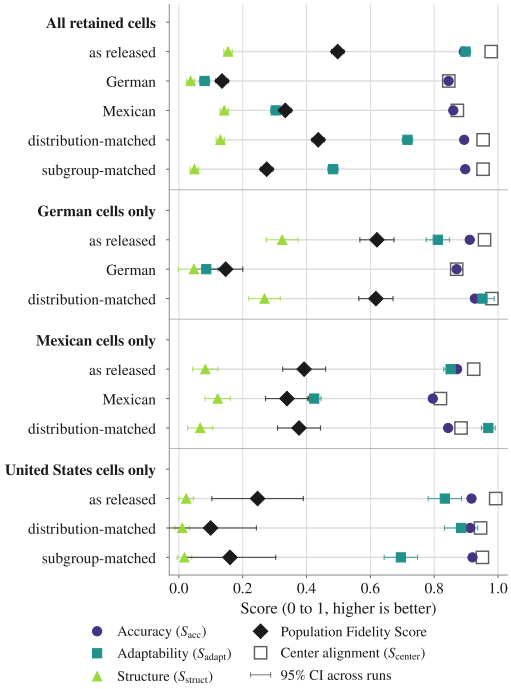
These results should be interpreted as transfer across datasets and population definitions: SubPOP is trained on United States demographic subgroups, whereas our evaluation compares demographic cells across five WVS countries. They therefore show that subgroup-level distributional training does not automatically transfer to the between-group structure of these data, rather than that subgroup matching generally fails to recover population structure.

Because the CultureLLM-style conditions target individual countries, Figure 16 also evaluates all conditions on German and Mexican cells separately. Within Germany, $\mathcal{F}_{\mathrm{de}}$ reduces both center alignment and PFS in six of eight combinations and compresses variation among German cells (median $A = 0.164$ versus 0.733 for the as-released model). Within Mexico, $\mathcal{F}_{\mathrm{mx}}$ raises center alignment by 0.103 on average, but PFS improves in only two combinations and declines in four; median A falls from 0.728 to 0.453.

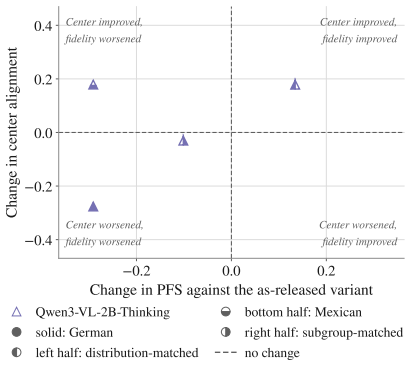
The additional distribution-trained conditions likewise show that changes in center alignment do not uniquely determine changes in PFS. Across the target-country evaluations, structure remains the limiting component in 77 of 80 combinations. These experiments involve only one backbone and substantially different training targets and data coverage, so they are intended to demonstrate how PFS distinguishes adaptation profiles rather than to establish comparative performance among fine-tuning methods.



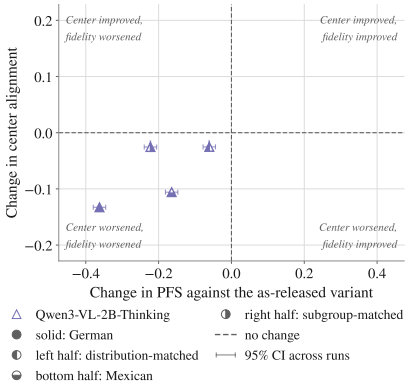
(a) Social trust, NTP: component profiles on all cells, German cells, and Mexican cells.



(b) Happiness, FA: component profiles on all cells, German cells, and Mexican cells.



(c) Social trust, NTP: change from as released, all cells.



(d) Happiness, FA: change from as released, all cells.

Figure 16: Population Fidelity and center alignment for five Qwen3-VL-2B-Thinking conditions: as released, $\mathcal{F}_{de}$, $\mathcal{F}_{mx}$, distribution-matched, and subgroup-matched. **(a,b)** Component profiles on all retained cells, German cells, and Mexican cells for social trust under NTP and happiness under FA, respectively. Diamonds show PFS and hollow squares center alignment. **(c,d)** Adapted minus as-released changes on all retained cells for the same two settings.