FairSynth: Instrument-Relative Evaluation and Calibration of LLM Survey Simulations for Measurement Fairness

Guangrui Fan[1]*, Dandan Liu[2], Qian Guo[1], Xueying Niu[1], Lihu Pan[1]

[1]School of Computer Science and Technology,
Taiyuan University of Science and Technology,
[2]Department of Media and Communication Studies, Universiti Malaya

fgr@tyust.edu.cn, s2134717@siswa.um.edu.my, 2022055@tyust.edu.cn,
xueyingniu@tyust.edu.cn, panlh@tyust.edu.cn

Abstract

Large language models (LLMs) are increasingly used as virtual survey respondents, but most evaluations still compare only per-item marginal response distributions. This can miss two failures that matter for computational social science: synthetic respondents may distort the correlation structure among attitudes, and errors may be uneven across countries or demographic groups. We introduce FAIRSYNTH, an instrument-relative framework for auditing and calibrating closed-ended, answer-only LLM survey simulations against existing anchor surveys. FairSynth evaluates marginal fidelity, Structural Attitude Fidelity (item–item correlation similarity), group representation error, flattening, and between-group sign reversals. It then applies a model-agnostic post-hoc logit-shift layer with shared group parameters, reducing representation-error disparities while preserving attitude structure. On World Values Survey Wave 7 (WVS7) and General Social Survey (GSS) 2022, FairSynth reduces worst-group representation error and between-group sign reversals while improving structural fidelity on the evaluated questionnaires. FairSynth is not a replacement for human surveys; it is a measurement-fairness toolkit for auditing, stress-testing, and calibrating synthetic survey simulations when anchor data are available.

1 Introduction

A computational social scientist may want to use an LLM to simulate a World Values Survey-style questionnaire for many countries and demographic strata before running robustness checks or designing a new study. A persona-conditioned model can appear plausible when judged only by each item's marginal distribution, yet still flatten lower-education groups, misrepresent Global South respondents, reverse between-group differences, or destroy the correlations among trust, religion, and values items that make survey data scientifically useful. These failures are not just statistical nuisances: they can change downstream conclusions about which groups differ and which attitudes cohere.

LLMs are increasingly used as virtual survey respondents in public opinion research (Argyle et al., 2023; Buskirk et al., 2025; Ziems et al., 2024). In "silicon sampling," models conditioned on demographic backstories answer survey questions as specific subgroups (Argyle et al., 2023; Sun et al., 2024), with applications spanning elections (Miranda and Balbi, 2025), climate opinion (Lee et al., 2024), and survey benchmarks (Zhao et al., 2025; Cao et al., 2025). Persona-conditioned LLMs can approximate some group-level distributions (Argyle et al., 2023; Tzachristas et al., 2025), but they also exhibit option-order and wording sensitivity (Dominguez-Olmedo et al., 2024; Bisbee et al., 2024), perform better for Western or high-resource contexts (Qu and Wang, 2024; Tao et al., 2024), and can flatten marginalised groups (Wang et al., 2025).

We target three concrete uses where synthetic surveys can be helpful but must be audited carefully: (i) *diagnosing misrepresentation*, by identifying groups for which a simulator has large representation error or between-group sign reversals; (ii) *controlled robustness checks*, by stress-testing survey analyses under calibrated synthetic populations; and (iii) *instrument-specific calibration*, by improving simulations of an existing questionnaire when a real anchor survey is available. FairSynth is not intended to replace the anchor survey. It is intended to audit LLM survey simulators, identify groups where they misrepresent an existing instrument, and support controlled robustness analyses under calibrated synthetic populations.

Despite rapid progress, three gaps limit principled use of LLM-based survey simulations. First,

*Corresponding author.

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 14992–15012
October 24-29, 2026 ©2026 Association for Computational Linguistics

item-wise metrics dominate: most evaluations focus on per-item distributions while ignoring structural attitude relationships central to survey-based social science (World Values Survey Association, 2020; Argyle et al., 2023; Cao et al., 2025). Second, fairness is often measured but not optimised: existing audits describe cross-group misrepresentation but rarely reduce it (Wang et al., 2025). Third, there is no standard calibration framework that jointly preserves survey structure and controls group-level representation error; fine-tuning can be expensive, unavailable for API-only models, and focused on independent item marginals (Cao et al., 2025).

We address these gaps with FAIRSYNTH, a framework for evaluating and post-hoc calibrating LLM-based survey simulators. We focus on *measurement fairness*: whether different groups are represented with comparable error relative to a fixed survey instrument, rather than how opinions ought to be distributed. We make three contributions. First, we formalise instrument-relative measurement fairness as parity in group-level representation error. Second, we introduce a multi-level evaluation suite combining marginal fidelity, Structural Attitude Fidelity (SAF), flattening, and between-group distortion. Third, we propose a lightweight, model-agnostic calibration layer that reduces worst-group representation error while preserving item–item attitude structure.

FairSynth assumes closed-ended survey items, answer-only prompting, access to option probabilities, and an existing anchor survey. Calibration is *instrument-relative*: it improves fidelity to a fixed questionnaire and snapshot. We do not claim that it forecasts new public opinion, replaces human surveys, guarantees normative fairness, or transfers automatically to unrelated instruments.

Relation to prior work. Appendix A situates FairSynth relative to silicon sampling, LLM survey simulation, cultural-bias audits, and distributional fine-tuning; our main distinction is to treat survey simulation as an instrument-relative measurement problem.

2 Method

We view an LLM-based survey simulator as a conditional distribution over closed-ended survey responses given respondent attributes. FairSynth has three components: (i) persona-conditioned, answer-only simulation; (ii) metrics for marginal fidelity, attitude structure, and group-level misrepresentation; and (iii) a post-hoc calibration layer that adjusts option probabilities without changing the base LLM.

2.1 Problem Setup

Let $x_i \in \mathcal{X}$ denote observed attributes of respondent i (e.g., age, gender, education, country), and let $y_{i,j} \in \mathcal{Y}_j$ denote the response of i to item $j \in J$, where $\mathcal{Y}_j = \{1, \dots, K_j\}$ is a finite set of options. We define groups $g \in \mathcal{G}$ as functions of the attributes, for example $g = \phi(x_i)$ where ϕ discretises age or education. Let D_g be the set of respondents with group label g, and let $D_{g,j} \subseteq D_g$ denote those with non-missing responses to item j. For each pair (g, j) we estimate the empirical human distribution

$$P_{g,j}(k) = \frac{1}{|D_{g,j}|} \sum_{i \in D_{g,j}} \mathbf{1}[y_{i,j} = k], \qquad (1)$$

and treat $\{P_{g,j}\}$ together with the induced inter-item correlations as the anchor measurement for that questionnaire and snapshot, rather than as timeless social truth.

A survey simulator is a conditional distribution $Q_\theta(y_{1:|J|} \mid x)$ induced by an LLM with a prompting scheme. Given attributes x_i and a prompt, the model produces a categorical distribution over $\mathcal{Y}_j$ for each item j, from which we sample synthetic responses $\hat{y}_{i,j}$. Synthetic group-wise distributions can be estimated from sampled labels ($\hat{P}_{g,j}$) or from probability-averaged outputs ($\bar{P}_{g,j}(k) = \frac{1}{|\hat{D}_g|} \sum_i q_{i,j}(k)$, where $q_{i,j}$ is the model's output distribution). We use $\bar{P}_{g,j}$ for marginal metrics (lower variance) and sampled labels for correlations.

2.2 Persona Construction and Prompting

Following Argyle et al. (2023); Tzachristas et al. (2025), we map attributes x_i to a textual persona π_i encoding demographics and, for guided personas, optional anchor attitudes. Each survey item is shown as a multiple-choice question with labelled response options. We randomise option order at generation time and map the selected label back to the original questionnaire scale before computing metrics, reducing simple position and label biases (Dominguez-Olmedo et al., 2024; Kabir et al., 2025). Guided-persona anchors are constructed only from pre-specified calibration attributes/items and never from held-out evaluation items or test responses. Prompt templates are in Appendix B.1.

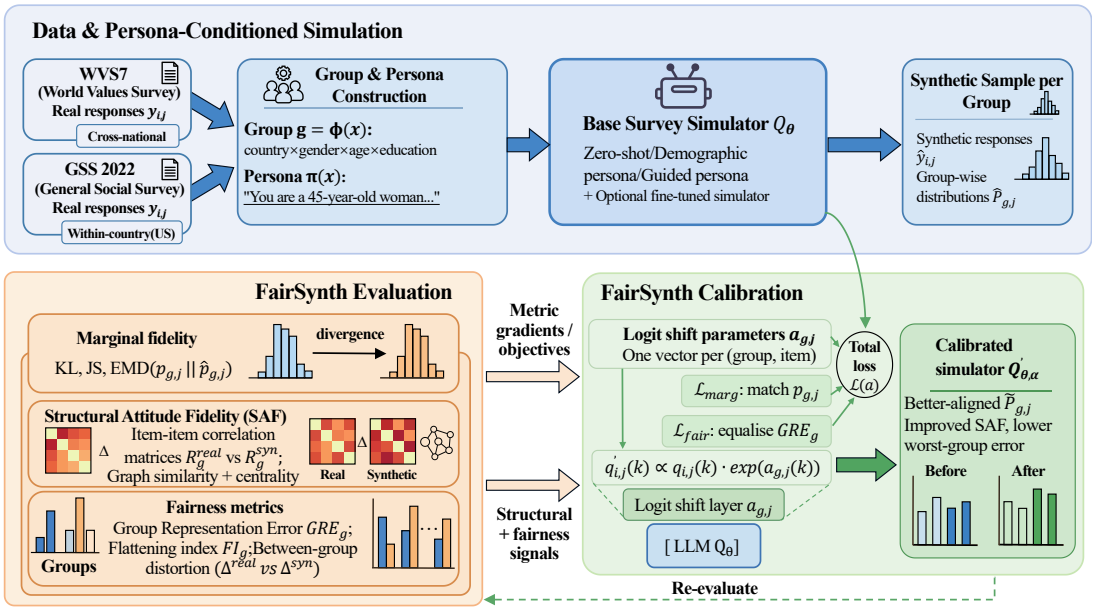


Figure 1: FairSynth overview. (A) Real survey data defines groups g and personas $\pi(x)$ conditioning simulator Q_θ. (B) Evaluation compares marginals (KL/EMD), structural fidelity (SAF), and fairness. (C) Calibration learns shared B_g yielding $a_{g,j}{=}\Phi_j B_g e_j$, enabling item generalisation.

2.3 Answer-only Constrained Decoding

FairSynth is designed for the closed-ended, answer-only setting rather than free-form survey dialogue. For every persona–item pair, the prompt ends with an instruction such as *"Respond with only the option label."* We decode at most one new token (`max_new_tokens=1`) and extract log-probabilities for the valid option labels only. The probabilities of these labels are then renormalised to obtain $q_{i,j} \in \Delta^{K_j-1}$. For items with many categories we use letter labels to avoid multi-token digit labels; implementation details are in Appendix F.5.

Because first-token multiple-choice scores can disagree with free-form answers (Wang et al., 2024), our claims apply only to constrained label-choice simulations. We renormalise probabilities over admissible option labels, never calibrate rationales or free-form text, and report interface diagnostics in Appendix F.5.

2.4 Base Survey Simulators

FairSynth treats simulators as black boxes. Q_θ may be zero-shot, persona-based (Argyle et al., 2023), or distributionally specialised (Cao et al., 2025). The only interface requirement is access to option probabilities $Q_\theta(\cdot \mid x_i, \pi_i)$ for the valid answer labels; gradients through the LLM are not required.

2.5 FairSynth Evaluation Metrics

We introduce a hierarchy of metrics for evaluating survey simulators: marginal fidelity, structural attitude fidelity, and group-level fairness and distortion. For ordinal items we encode responses as integers $s_{i,j} \in \{1, \ldots, K_j\}$ and use the same encoding for synthetic responses $\hat{s}_{i,j}$. Group-wise human and synthetic marginals are $P_{g,j}$ and $\hat{P}_{g,j}$.

Marginal fidelity. For each (g, j) we compute standard distributional distances between $P_{g,j}$ and $\hat{P}_{g,j}$ such as Kullback–Leibler divergence and Earth Mover's Distance (EMD). As a per-group marginal fidelity measure we report

$$\mathrm{KL}_g = \frac{1}{|J|} \sum_j \mathrm{KL}\big(P_{g,j} \,\|\, \hat{P}_{g,j}\big), \tag{2}$$

and we aggregate EMD across items via the group representation error GRE_g defined below. For ordinal items, EMD is computed on the ordered support $\{1, \ldots, K_j\}$ with unit spacing, so it is sensitive to the distance between response categories. These metrics are widely used in prior work on LLM-based survey simulation (Argyle et al., 2023; Lee et al., 2024; Tzachristas et al., 2025; Miranda and Balbi, 2025; Cao et al., 2025).

Structural attitude fidelity. Marginals do not capture how items relate to each other. For each group g we estimate correlation matrices $R_g^{\text{real}}, R_g^{\text{syn}} \in \mathbb{R}^{|J|\times|J|}$, where

$$r_{g,jk}^{\text{real}} = \mathrm{corr}(s_{i,j}, s_{i,k})_{i \in D_g}, \tag{3}$$

$$r_{g,jk}^{\text{syn}} = \mathrm{corr}(\hat{s}_{i,j}, \hat{s}_{i,k})_{i \in \hat{D}_g}. \tag{4}$$

We summarise structural similarity via three complementary metrics: (i) the Frobenius norm $\Delta_g^{\mathrm{F}} = \|R_g^{\mathrm{real}} - R_g^{\mathrm{syn}}\|_F$, (ii) the Spearman correlation ρ_g^{deg} between node degrees in the two item graphs (where degree of item j is $\sum_{k \neq j} |r_{jk}|$), and (iii) a simple entry-wise correlation ρ_g^{corr} defined as the Spearman correlation between the upper-triangular entries of R_g^{real} and R_g^{syn} (i.e., all $\binom{|J|}{2}$ pairwise item correlations). We combine the first two into a normalised Structural Attitude Fidelity score

$$\mathrm{SAF}_g = 1 - \alpha\, \widetilde{\Delta_g^{\mathrm{F}}} - (1-\alpha)\big(1 - \rho_g^{\mathrm{deg}}\big), \quad (5)$$

where $\widetilde{\Delta_g^{\mathrm{F}}}$ rescales Δ_g^{F} to $[0,1]$ across groups and $\alpha \in [0,1]$ trades off matrix similarity and centrality preservation. Intuitively, SAF is high when pairwise item associations and central attitude items are preserved. Because SAF is composite, Appendix B.4 reports raw Frobenius distance, degree correlation, and entry-wise correlation; all move in the same direction as the aggregate score.

Group representation error and flattening. We define

$$\mathrm{GRE}_g = \frac{1}{|J|} \sum_j \mathrm{EMD}\big(P_{g,j}, \hat{P}_{g,j}\big), \quad (6)$$

which measures how well the simulator represents group g on average. Inspired by evidence that LLMs can "flatten" identity groups (Wang et al., 2025), we also track the ratio of synthetic to human variance:

$$\mathrm{Var}_{g,j}^{\mathrm{real}} = \mathrm{Var}\big(s_{i,j}\big)_{i \in D_g}, \quad (7)$$

$$\mathrm{Var}_{g,j}^{\mathrm{syn}} = \mathrm{Var}\big(\hat{s}_{i,j}\big)_{i \in \hat{D}_g}, \quad (8)$$

$$\mathrm{FI}_g = \frac{1}{|J|} \sum_j \frac{\mathrm{Var}_{g,j}^{\mathrm{syn}}}{\mathrm{Var}_{g,j}^{\mathrm{real}} + \epsilon}, \quad (9)$$

with a small $\epsilon > 0$ for stability. Values $\mathrm{FI}_g < 1$ indicate systematic variance shrinkage (flattening) for group g, whereas $\mathrm{FI}_g > 1$ indicates that the simulator produces more dispersed responses than the survey for group g. FI is a coarse, relative indicator of variance shrinkage; we focus on cross-model comparisons rather than absolute values.

Our metrics target measurement fairness—equalising representation error across groups relative to an anchor survey—rather than normative fairness or downstream decision-making. GRE thus measures relative misrepresentation of group g, and calibration reduces disparities in GRE rather than equalising opinions themselves.

Between-group distortion. To assess how well relative differences between groups are preserved, we compute, for each pair (g, g') and item j, the real and synthetic mean differences

$$\Delta_{g,g',j}^{\mathrm{real}} = \mu_{g,j}^{\mathrm{real}} - \mu_{g',j}^{\mathrm{real}}, \quad (10)$$

$$\Delta_{g,g',j}^{\mathrm{syn}} = \mu_{g,j}^{\mathrm{syn}} - \mu_{g',j}^{\mathrm{syn}}, \quad (11)$$

where $\mu_{g,j}^{\mathrm{real}}$ and $\mu_{g,j}^{\mathrm{syn}}$ are means of $s_{i,j}$ and $\hat{s}_{i,j}$ respectively. We compute these quantities over all ordered triples (g, g', j) such that both groups have sufficient support for item j (at least a fixed minimum number of non-missing responses), and we report (i) the Pearson correlation between $\Delta_{g,g',j}^{\mathrm{real}}$ and $\Delta_{g,g',j}^{\mathrm{syn}}$ across all such triples and (ii) the proportion of triples for which $\mathrm{sign}(\Delta_{g,g',j}^{\mathrm{real}}) \neq \mathrm{sign}(\Delta_{g,g',j}^{\mathrm{syn}})$. High sign-flip rates indicate severe between-group representation distortions. Main tables report all-triple sign-flip rates; because near-zero group differences can make signs unstable, Appendix F.4 analyses dead-zone thresholds that exclude triples with small $|\Delta_{g,g',j}^{\mathrm{real}}|$.

2.6 FairSynth Calibration

FAIRSYNTH adds a post-hoc calibration layer that adjusts the categorical distribution of a fixed base simulator Q_θ. The layer is fit against an anchor survey, but it is deliberately constrained: it learns a small number of group-level parameters shared across items rather than independent corrections for every group–item pair. The calibrator is black-box and requires only the base option probabilities $q_{i,j} \in \Delta^{K_j - 1}$ for each respondent persona i and item j.

2.6.1 Shared logit-shift parameterisation

For each group g we learn a matrix $B_g \in \mathbb{R}^{d \times r}$. For each survey item j we precompute a fixed item embedding $e_j \in \mathbb{R}^r$ and a fixed response-basis matrix $\Phi_j \in \mathbb{R}^{K_j \times d}$ (defined below). The calibrated logit shift vector for group g and item j is

$$a_{g,j} \;=\; \Phi_j\, B_g\, e_j \in \mathbb{R}^{K_j}. \quad (12)$$

This replaces independent per-(g, j) parameters with shared group parameters B_g that generalise across items via e_j. Because adding a constant to $a_{g,j}$ does not change a softmax, we construct Φ_j to be mean-zero across categories, yielding identifiable shifts. This shared parameterisation enables within-instrument transfer to held-out items because the same group parameters B_g modulate all items via fixed embeddings e_j.

Given base probabilities $q_{i,j}$ from Q_θ for a respondent $i \in \hat{D}_g$ in group g, we define calibrated probabilities:

$$q'_{i,j}(k) = \frac{q_{i,j}(k)\exp(a_{g,j}(k))}{\sum_{k'=1}^{K_j} q_{i,j}(k')\exp(a_{g,j}(k'))}. \quad (13)$$

This transformation is a distributional wrapper around the simulator: it changes only the probability of each admissible answer label and leaves the underlying LLM, persona text, and option wording unchanged. Because the generation protocol contains no rationale, there is no separate reasoning string that can become inconsistent with the calibrated answer.

We form calibrated group-level distributions by averaging over synthetic respondents:

$$\tilde{P}_{g,j}(k;B) = \frac{1}{|\hat{D}_g|}\sum_{i\in\hat{D}_g} q'_{i,j}(k). \quad (14)$$

Response basis and item embeddings. For ordinal items, Φ_j is a fixed mean-zero basis capturing directional shifts and response extremity; for binary items it reduces to one effective degree of freedom. Item embeddings e_j are fixed principal-coordinate embeddings derived from the base simulator's pooled synthetic item-correlation matrix. Details are in Appendix B.2.

2.6.2 Multi-objective calibration objective

Optimising marginals alone degrades structure; optimising structure alone leaves fairness disparities. We therefore learn $B = \{B_g\}_{g\in\mathcal{G}}$ by minimising a multi-objective loss:

$$\mathcal{L}(B) = \mathcal{L}_{\text{marg}}(B) + \lambda_{\text{struct}}\mathcal{L}_{\text{struct}}(B) + \lambda_{\text{fair}}\mathcal{L}_{\text{fair}}(B) + \lambda_{\text{reg}}\sum_g \|B_g\|_F^2, \quad (15)$$

where the terms correspond to marginal alignment, structural preservation, fairness, and regularisation. The marginal term aligns calibrated marginals with human marginals:

$$\mathcal{L}_{\text{marg}}(B) = \sum_{g,j}\text{KL}\Big(P_{g,j}\,\|\,\tilde{P}_{g,j}(\cdot;B)\Big). \quad (16)$$

This is analogous to objectives used in distributional fine-tuning (Cao et al., 2025) but operates on top of a fixed simulator. For ordinal items we compute $\text{EMD}(P_{g,j}, \tilde{P}_{g,j}(\cdot;B))$ on the ordered support $\{1,\dots,K_j\}$ via the ℓ_1 distance between cumulative distribution functions, using a smooth approximation to the absolute value to obtain stable gradients (details in Appendix C.4). The structural term penalises changes to the simulator's pre-calibration correlation structure:

$$\mathcal{L}_{\text{struct}}(B) = \sum_g \big\|R_g^{\text{cal}}(B) - R_g^{\text{syn}}\big\|_F^2, \quad (17)$$

where $R_g^{\text{cal}}(B)$ is the calibrated item–item correlation matrix (computed from soft expected scores $m_{i,j} = \sum_k k \cdot q'_{i,j}(k)$ during training for differentiability) and R_g^{syn} is the base simulator's matrix for group g. We deliberately anchor this term to the simulator's own pre-calibration structure R_g^{syn}, rather than directly to the human correlations R_g^{real}, to avoid degenerate solutions that overfit noisy survey covariances or entirely overwrite the LLM's internal attitude graph. Intuitively, anchoring to R_g^{syn} makes calibration "local" in attitude-structure space: we adjust the simulator's outputs while staying close to its internal belief network, rather than forcing it to match potentially noisy empirical covariances. Empirically (Appendix E.4), directly matching R_g^{real} can overfit noisy empirical covariances for small groups and substantially increase worst-group GRE, whereas anchoring to R_g^{syn} yields more stable trade-offs between marginal fidelity, structure, and fairness.

The fairness term equalises representation error across groups:

$$\mathcal{L}_{\text{fair}}(B) = \text{Var}_{g\in\mathcal{G}}\big(\text{GRE}'_g(B)\big), \quad (18)$$

where $\text{GRE}'_g(B) = \frac{1}{|J|}\sum_j \text{EMD}\Big(P_{g,j}, \tilde{P}_{g,j}(\cdot;B)\Big)$. The variance-based fairness term encourages parity in GRE while allowing overall error reduction. Alternative objectives (max–min, weighted GRE) are compared in Appendix B.6; we adopt variance-based loss as it provides balanced trade-offs.

2.6.3 Optimisation and use

We optimise B with mini-batch gradient methods, keeping Q_θ and the fixed quantities (Φ_j, e_j) unchanged. Each step samples a mini-batch of groups and synthetic respondents, computes calibrated probabilities via Equations 12–13, forms mini-batch estimates of $\tilde{P}_{g,j}$ and R_g^{cal}, and updates B. Because the shared parameterisation reduces the number of learned parameters from $O(|\mathcal{G}||J|\bar{K})$ to $O(|\mathcal{G}|rd)$ (e.g., from ∼11,000 to ∼2,500 for

Dataset	# items	# respondents	# groups
WVS7	26	88,386	156
GSS 2022	19	3,441	17

Table 1: Survey subsets used for evaluation. WVS7 groups are 156 merged demographic strata covering all 64 countries, based on country × gender × age band × education with rare cells merged.

WVS7 with $r{=}8, d{=}2$), optimisation is fast and less prone to overfitting small groups. Algorithmic details and pseudocode are provided in Appendix B.5. At inference time we wrap Q_θ with the learned B, yielding a calibrated simulator $Q'_{\theta,B}$ that we evaluate with the same metrics.

3 Experimental Setup

3.1 Datasets and Group Construction

We use World Values Survey Wave 7 (WVS7) (World Values Survey Association, 2020) for cross-national evaluation and General Social Survey 2022 (GSS 2022) (NORC at the University of Chicago, 2023) for within-country intersectional analysis. For WVS7, we construct 156 merged demographic strata covering all 64 countries, based on country, gender, age band, and education level, with rare cells merged to satisfy minimum support. For GSS, we construct race × gender × party strata with the same minimum-support rule. Table 1 shows basic statistics; details in Appendix C.1. WVS7 and GSS 2022 are historical snapshots; throughout, performance means fidelity to those measured snapshots rather than prediction of current or future opinion. Both datasets provide survey weights, but the main experiments use unweighted responses and treat the realised sample as the population of interest. Appendix F.1 shows that the main FairSynth gains are stable under weighted evaluation.

3.2 Base Survey Simulators and Baselines

We instantiate five base simulators Q_θ spanning common design choices: (i) zero-shot prompting with no persona (lower bound); (ii) demographic persona conditioning on age, gender, location, and education (Argyle et al., 2023; Sun et al., 2024); (iii) guided persona adding anchor attitudes such as political orientation (Tzachristas et al., 2025); (iv) a distributionally specialised simulator fine-tuned to match empirical marginals (Cao et al., 2025); and (v) a non-LLM statistical baseline using multinomial logistic regression (Miranda and Balbi, 2025). FairSynth treats all simulators as black boxes exposing categorical distributions over $\mathcal{Y}_j$; model-specific details are in Appendix C.3. Unless otherwise noted, main-text results use Qwen2.5-72B as the underlying LLM for all prompt-based simulators; we report cross-model comparisons with other open-source models (Qwen2.5-7B, Llama-3.1-8B) in Appendix C.3, finding qualitatively similar trends. Guided-persona anchors are restricted to pre-specified calibration attributes/items and never include held-out evaluation items or held-out test responses. We also include a non-deployable histogram oracle that samples from empirical $P_{g,j}$; it is purely diagnostic and illustrates how near-perfect marginal alignment can still destroy item–item structure.

We include a *temperature scaling* baseline that optimises per-(g, j) temperatures to minimise KL divergence on J_{calib}. It preserves within-persona structure for SAF computation but has no structural or fairness term and cannot generalise to held-out items.

3.3 Calibration Configurations

We split respondents into train/validation/test sets at the individual level (60/20/20) within each dataset, ensuring that no respondent appears in more than one split and that all 156 WVS7 groups (17 for GSS) are represented in each split. Calibration parameters B_g are learned using only training respondents: empirical human marginals $P_{g,j}$ in $\mathcal{L}_{\text{marg}}$ are estimated on the training split, and $\text{GRE}'_g(B)$ in $\mathcal{L}_{\text{fair}}$ is computed with respect to those same training marginals. Hyperparameters and early stopping are selected on the validation split. All metrics reported in Section 4 (KL, SAF, GRE, FI, sign flips) are computed on the held-out test respondents. The shared parameterisation yields $\sim$2,500 learned parameters for WVS7 ($156 \times r{=}8 \times d{=}2$) and $\sim$270 for GSS, substantially fewer than per-(g, j) alternatives. We initialise $B_g = 0$, optimise with Adam, and enforce Frobenius norm constraints to prevent overfitting. Details in Appendix C.4.

Because calibration uses anchor data, overfitting is the main validity threat. We therefore evaluate only on held-out respondents, use a shared B_g rather than independent per-(g, j) shifts, regularise $\|B_g\|_F$, and test held-out items whose human marginals are never used during calibration. We also include an anchoring ablation (Appendix E.4) showing that directly forcing calibrated correlations toward R_g^{real} increases worst-group GRE,

whereas anchoring structural changes locally to R_g^{syn} gives a more stable fairness–structure trade-off.

3.4 Evaluation Protocol

We evaluate simulators in base (raw Q_θ) and calibrated (with learned B_g) regimes. Metrics include KL/EMD for marginals, SAF for structure, GRE/FI for group fairness, and sign-flip rate for between-group distortion (Appendix B.3). We also partition items into calibration (J_{calib}, 2/3) and held-out (J_{held}, 1/3) sets to test item-level generalisation. Our analysis addresses: (1) baseline simulator fidelity and error dispersion; (2) FairSynth improvements in GRE, SAF, and between-group distortion; and (3) robustness across simulators and settings.

For item-level generalisation, the held-out experiment trains the human-supervised losses only on J_{calib} and reports KL/GRE/SAF separately on J_{calib} and J_{held} using held-out test respondents. Because B_g is shared across items through fixed embeddings e_j, calibration can still affect held-out items whose human marginals are never used for fitting. Details are in Appendix E.3.

4 Results

4.1 Marginal Fidelity Alone Misses Attitude Structure

Table 2 reports aggregate performance on WVS7 and GSS. Zero-shot prompting performs worst; adding demographic and guided personas progressively improves the uncalibrated simulator. However, the baselines designed to match marginals reveal a central failure mode. The histogram oracle drives KL/GRE near zero but damages structure (SAF drops from 0.76 to 0.58 on WVS7), because it samples each item independently. Temperature scaling improves KL but degrades SAF from 0.76 to 0.68 because it learns independent (g, j) adjustments with no structural or fairness constraint. FairSynth provides the best balanced outcome: on WVS7 it reduces KL by 47% (0.17→0.09), reduces GRE by 25% (0.16→0.12), and improves SAF from 0.76 to 0.82; the same pattern holds on GSS. These results support the main premise of the paper: per-item marginal alignment is insufficient for survey-style inference.

	WVS7			GSS		
Simulator	KL↓	SAF↑	GRE↓	KL↓	SAF↑	GRE↓
Tabular baseline	.29	.52	.24	.26	.55	.21
Zero-shot LLM	.38	.57	.28	.32	.59	.24
Demog. persona	.25	.68	.20	.22	.71	.17
Guided persona	.17	.76	.16	.15	.78	.14
+ hist. oracle	.02	.58	.02	.02	.60	.02
+ temp. scal.	**.07**	.68	.13	**.06**	.70	.11
+ FairSynth	.09	**.82**	**.12**	.08	**.84**	**.10**

Table 2: Aggregate performance (mean across 3 seeds; SDs in Appendix D.1). Bold marks the best non-oracle scalar value. Green : selected deployable method with the best overall trade-off; the histogram oracle is diagnostic and excluded from deployable-method ranking. Yellow : strongest uncalibrated prompt-based simulator.

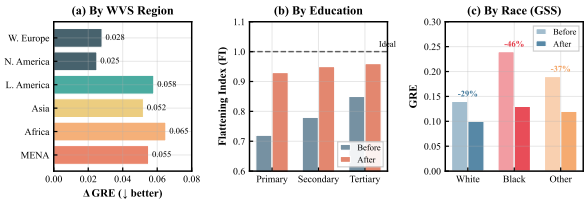


Figure 2: Heterogeneous FairSynth gains. (a) By region: largest GRE reduction for Africa/Latin America. (b) By education: FI improves most for low education. (c) By race: Black/Other groups see 40–50% GRE reduction vs. smaller gains for White groups.

	WVS7				GSS			
Simulator	Std	Max	FI	Flip	Std	Max	FI	Flip
Guided persona	.054	.35	.78	.11	.048	.32	.83	.09
+ hist. orc.	.008	.04	1.0	.20	.007	.03	1.0	.18
+ temp. scal.	.046	.31	.85	.09	.041	.28	.88	.08
+ FairSynth	**.028**	**.19**	**.94**	**.05**	**.025**	**.17**	**.96**	**.04**
Specialised	.085	.42	.69	.10	.072	.35	.75	.08
+ FairSynth	**.032**	**.18**	**.92**	**.04**	**.028**	**.16**	**.94**	**.03**

Table 3: Fairness metrics. Std/Max: standard deviation and maximum of GRE; FI: flattening index (ideal≈1); Flip: between-group sign-flip rate. Green/bold marks the selected deployable calibrated method; the histogram oracle is diagnostic and excluded from deployable-method ranking.

4.2 FairSynth Reduces Worst-Group Error and Sign Flips

Table 3 reports fairness-oriented diagnostics. The uncalibrated guided persona simulator has substantial error dispersion (Std(GRE)=0.054 and max GRE=0.35 on WVS7) and pronounced flattening (FI=0.78). The histogram oracle obtains artificially low GRE by sampling from empirical per-item marginals, but it is diagnostic only; its sign-flip value is sensitive to finite-sample independent item sampling and should not be interpreted as a deployable fairness result. FairSynth reduces Std(GRE)

from .054 to .028, reduces worst-group GRE from .35 to .19, moves FI toward one, and cuts sign flips from 11% to 5% on WVS7. Similar patterns hold on GSS and for the specialised simulator.

The gains are heterogeneous in the desired direction: groups with larger initial errors tend to receive larger GRE reductions. Region-, education-, and race-level breakdowns are reported in Figure 2 and Appendix E.1.

4.3 Ablation: Each Objective Prevents a Failure Mode

Table 4 reports a loss-component ablation for the guided-persona WVS7 setting. Marginal-only calibration achieves the lowest KL, but damages attitude structure (SAF 0.65) and leaves worst-group error high. Adding fairness without structure yields the lowest GRE dispersion, but further harms SAF. Adding structure without fairness restores SAF, but leaves error dispersion and worst-group error too high. Full FairSynth provides the best balance across KL, SAF, and worst-group robustness, although marginal+fair achieves lower GRE dispersion at the cost of structural fidelity.

Variant (WVS7)	KL↓	SAF↑	Std↓	Max↓
Guided (base)	.17±.05	.76±.07	.054±.02	.35
Marginal only	**.05**±.02	.65±.09	.048±.02	.31
Marg. + fair	.07±.02	.63±.10	**.020**±.01	**.18**
Marg. + struct	.10±.03	.79±.06	.038±.01	.28
Full FairSynth	.09±.03	**.82**±.05	.028±.01	.19

Table 4: Loss-component ablation on WVS7. Bold marks the best value in each column; the green row marks the selected full objective, which provides the best balance across marginal fidelity, structural fidelity, and worst-group robustness.

4.4 Structural Fidelity and Held-out Items

Across all WVS7 groups, SAF increases from 0.76 to 0.82 under FairSynth; on GSS, SAF rises from 0.78 to 0.84. Appendix B.4 shows that this is not an artefact of the composite SAF score: raw Frobenius distance, degree correlation, and entry-wise correlation all improve. Qualitative heatmaps and item-graph visualisations appear in Appendix E.6.

Table 5 tests whether calibration merely memorises fitted item marginals. Human-supervised losses use only J_{calib}; no human marginals or correlations from J_{held} are used. FairSynth still improves held-out WVS7 items (KL 0.18→0.11, GRE 0.17→0.13, SAF 0.74→0.79), with similar GSS gains. This supports within-instrument transfer, not arbitrary cross-instrument generalisation.

Data	Items	KL↓	SAF↑	GRE↓
WVS7	J_{calib} base	.17±.05	.76±.07	.16±.04
WVS7	J_{calib} FS	.09±.03	.83±.05	.12±.03
WVS7	J_{held} base	.18±.06	.74±.08	.17±.05
WVS7	J_{held} FS	.11±.04	.79±.06	.13±.03
GSS	J_{held} base	.16±.05	.76±.07	.15±.04
GSS	J_{held} FS	.09±.03	.80±.05	.12±.03

Table 5: Within-instrument item generalisation. FS denotes FairSynth. Human-supervised losses are fit only on J_{calib}; J_{held} results are evaluated after training. Full calibration/held-out results are in Appendix E.3.

4.5 Between-group Differences

FairSynth increases the correlation between real and synthetic group-pair differences from 0.79 to 0.87 on WVS7 and from 0.81 to 0.89 on GSS. It also reduces sign-flip rates from 11% to 5% on WVS7 and from 9% to 4% on GSS. Sign-flip improvements are stable under different dead-zone thresholds (Appendix F.4). These results are important because sign flips correspond to synthetic data reversing the direction of a group comparison for an item, which is a severe failure for survey-based analysis.

5 Discussion and Conclusion

FairSynth is most useful when an established questionnaire exists and researchers want to audit, stress-test, or calibrate synthetic survey simulations relative to that instrument. Our results show why this instrument-relative framing matters: marginal-only methods can look strong by KL/GRE while damaging attitude structure or reversing between-group differences. FairSynth reduces these failures by constraining calibration through shared group parameters and by balancing marginal alignment, structural preservation, and GRE parity.

The framework should not be read as evidence that LLMs replace human samples or forecast current opinion. Anchor surveys encode sampling choices, nonresponse patterns, and possible measurement-invariance violations; calibration improves fidelity to observed measurements, not to an unobserved social truth. We introduced FairSynth as an evaluation and calibration framework for closed-ended LLM survey simulations; on WVS7 and GSS 2022, it reduces worst-group error and sign reversals while improving structural fidelity, with the largest gains for groups that the base simulator misrepresents most.

Limitations

FairSynth assumes access to high-quality anchor surveys; sparsely sampled groups cannot be calibrated reliably and may remain poorly represented. Our experiments focus on English prompt renderings of 19–26 closed-ended items from WVS7 and GSS 2022. Calibration and evaluation occur within the same questionnaires, with held-out item tests showing only partial within-instrument generalisation. Transfer to richer instruments, open-ended responses, longitudinal designs, or entirely new questionnaires remains untested.

Our fairness objective is measurement-oriented. It equalises representation error relative to existing survey responses, but it does not guarantee normative fairness, does not correct biases in the anchor surveys, and does not address whether items are measurement-invariant across countries or demographic groups. FairSynth operates at the level of observed responses; if the same response option reflects different latent constructs across groups, calibration may reinforce instrument-specific artefacts.

The method also requires access to token-level option probabilities. This is available for many LLM APIs and open-source inference stacks, but not all deployment settings expose it. The diagnostic protocol in Appendix F.5 helps audit this interface, but it does not remove the dependency on well-behaved answer-label probabilities. Because our protocol is answer-only and single-token, FairSynth does not address free-form rationales, multi-turn interviews, or open-ended survey responses. Finally, although the shared parameterisation reduces overfitting compared with per-(g, j) shifts, small cells can still yield noisy calibration targets; hierarchical extensions that pool information across similar groups are an important next step.

Ethical Considerations

This work uses publicly available, institutionally reviewed survey datasets (WVS7, GSS 2022); we collect no new human data. FairSynth is designed to audit and calibrate synthetic survey simulations, not replace human participants. We caution against using calibrated synthetic data in high-stakes contexts, forecasting settings, or policy decisions without new validation against real responses. Our fairness objectives are measurement-oriented and do not guarantee normative fairness or address biases in anchor surveys. Calibrated respondents may still misrepresent undersampled groups.

AI Writing Disclosure

AI assistants were used to support language editing and prose organisation. All scientific claims, experimental design, data analysis, and interpretation remain the responsibility of the authors. AI-generated suggestions were reviewed, edited, and verified by the authors before incorporation. No AI-generated content was incorporated without author review.

References

Badr AlKhamissi, Muhammad ElNokrashy, Mai AlKhamissi, and Mona Diab. 2024. Investigating cultural alignment of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL), Long Papers*, pages 12404–12422. Association for Computational Linguistics.

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

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

Trent D Buskirk, Florian Keusch, Leah von der Heyde, and Adam Eck. 2025. More parameters than populations: A systematic literature review of large language models within survey research. *arXiv preprint arXiv:2509.03391*.

Yong Cao, Haijiang Liu, Arnav Arora, Isabelle Augenstein, Paul Röttger, and Daniel Hershcovich. 2025. 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)*, pages 3141–3154.

Chris Chapman. 2025. Synthetic survey data? it's not data. https://quantuxblog.com/synthetic-survey-data-its-not-data. Blog post.

Suhyun Cho, Jaeyun Kim, and Jang Hyun Kim. 2024. Llm-based doppelgänger models: Leveraging synthetic data for human-like responses in survey simulations. *IEEE Access*, 12:178917–178927.

Ricardo Dominguez-Olmedo, Moritz Hardt, and Celestine Mendler-Dünner. 2024. Questioning the survey responses of large language models. In *Advances in Neural Information Processing Systems*. NeurIPS 2024.

Chengpiao Huang, Yuhang Wu, and Kaizheng Wang. 2025. How many human survey respondents is a large language model worth? an uncertainty quantification perspective. *arXiv preprint arXiv:2502.17773*.

Shapeng Jiang, Lijia Wei, and Chen Zhang. 2024. Donald trumps in the virtual polls: Simulating and predicting public opinions in surveys using large language models. *arXiv preprint arXiv:2411.01582*.

Mohsinul Kabir, Ajwad Abrar, and Sophia Ananiadou. 2025. Break the checkbox: Challenging closed-style evaluations of cultural alignment in LLMs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*.

Sanguk Lee, Tai-Quan Peng, Matthew H. Goldberg, Seth A. Rosenthal, John E. Kotcher, Edward W. Maibach, and Anthony Leiserowitz. 2024. Can large language models estimate public opinion about global warming? an empirical assessment of algorithmic fidelity and bias. *PLOS Climate*, 3(8):e0000429.

Ted K. Mburu, Kangxuan Rong, Campbell J. McColley, and Alexandra Werth. 2025. Methodological foundations for ai-driven survey question generation. *arXiv preprint arXiv:2505.01150*.

Fernando Miranda and Pedro Paulo Balbi. 2025. Simulating public opinion: Comparing distributional and individual-level predictions from LLMs and random forests. *Entropy*, 27(9):923.

Ha Nguyen, Victoria Nguyen, Saríah López-Fierro, Sara Ludovise, and Rossella Santagata. 2024. Simulating climate change discussion with large language models: Considerations for science communication at scale. In *Proceedings of the Eleventh ACM Conference on Learning at Scale*, pages 28–38.

NORC at the University of Chicago. 2023. General social survey 2022 cross-section. `https://gss.norc.org/`. GSS 2022 codebook and dataset.

Yao Qu and Jue Wang. 2024. Performance and biases of large language models in public opinion simulation. *Humanities and Social Sciences Communications*, 11(1095).

Seungjong Sun, Eungu Lee, Dongyan Nan, Xiangying Zhao, Wonbyung Lee, Bernard J. Jansen, and Jang Hyun Kim. 2024. Random silicon sampling: Simulating human sub-population opinion using a large language model based on group-level demographic information. *arXiv preprint arXiv:2402.18144*.

Yan Tao, Olga Viberg, Ryan S. Baker, and René F. Kizilcec. 2024. Cultural bias and cultural alignment of large language models. *PNAS Nexus*, 3(9):pgae346.

Ioannis Tzachristas, Santhanakrishnan Narayanan, and Constantinos Antoniou. 2025. Guided persona-based AI surveys: Can we replicate personal mobility preferences at scale using LLMs? *arXiv preprint arXiv:2501.13955*.

Sebastián Valenzuela, Stephan Winter, and Sebastián Rivera. 2025. Using large language models for survey research in communication: opportunities and challenges. *Communication and Change*, 1(1):14.

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

Xinpeng Wang, Bolei Ma, Chengzhi Hu, Leon Weber-Genzel, Paul Röttger, Frauke Kreuter, Dirk Hovy, and Barbara Plank. 2024. "my answer is C": First-token probabilities do not match text answers in instruction-tuned language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 7407–7416, Bangkok, Thailand. Association for Computational Linguistics.

World Values Survey Association. 2020. World values survey, wave 7 (2017–2020). `https://www.worldvaluessurvey.org/wvs.jsp`. Wave 7 data release.

Jianpeng Zhao, Chenyu Yuan, Weiming Luo, Haoling Xie, Guangwei Zhang, Steven Jige Quan, Zixuan Yuan, Pengyang Wang, and Denghui Zhang. 2025. Large language models as virtual survey respondents: Evaluating sociodemographic response generation. *arXiv preprint arXiv:2509.06337*.

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

A Extended Related Work

LLM-based survey simulation. The use of LLMs as virtual survey respondents has grown rapidly. Argyle et al. (2023) introduced "silicon sampling," showing that GPT-3 can approximate U.S. political attitudes when conditioned on detailed backstories. Subsequent work extended this paradigm to climate opinion (Lee et al., 2024; Nguyen et al., 2024), mobility preferences (Tzachristas et al., 2025), and electoral forecasting (Miranda and Balbi, 2025; Jiang et al., 2024). Benchmarks such as LLM-S^3 (Zhao et al., 2025) and fine-tuning approaches (Cao et al., 2025) aim to systematise evaluation and improve distributional fidelity. However, these efforts largely focus on per-item marginals and do not explicitly address structural or fairness concerns.

Pitfalls and biases. A parallel literature documents limitations. Dominguez-Olmedo et al. (2024) show that LLM survey responses exhibit strong option-order and label biases; Kabir et al. (2025) argue that closed-form questionnaires can systematically misrepresent cultural values; Bisbee et al. (2024) find that synthetic replacements are highly sensitive to wording. Cross-national studies reveal that LLM simulations work better for Western, high-resource contexts (Qu and Wang, 2024; Tao et al., 2024; Zhao et al., 2025), and Wang et al. (2025) show that LLMs can flatten marginalised identity groups. Huang et al. (2025) provide uncertainty quantification for synthetic surveys, and practitioners warn against overinterpretation (Chapman, 2025; Buskirk et al., 2025; Valenzuela et al., 2025; Mburu et al., 2025).

Fairness in generative models. Fairness constraints have been studied extensively in discriminative models but less so in generative settings. Recent work on text generation examines demographic biases in language models and proposes debiasing techniques, though these typically target toxicity or stereotype rather than survey fidelity. FairSynth adapts the idea of group-conditional error parity to the survey simulation setting, treating fairness as equalising representation error across demographic groups.

Structural attitude modelling. Survey methodologists have long emphasised that attitudes form structured networks, not isolated items (World Values Survey Association, 2020; AlKhamissi et al., 2024; Cho et al., 2024). Cross-cultural psychology uses correlation and factor structures to characterise value dimensions. FairSynth operationalises this insight via the Structural Attitude Fidelity metric, which compares item–item correlation matrices between human and synthetic data.

B Additional Method Details

B.1 Persona Templates

In all experiments we instantiate personas as short role-play descriptions derived from a fixed set of attributes (age, gender, education, country/location).

Demographic persona template (WVS7).

```
You are a [AGE_DESC] [GENDER] from
[COUNTRY] [EDUCATION_DESC].
Answer this survey question from your
perspective.
Respond with only the number.
```

Guided persona template (WVS7). Extends the demographic template with attitude anchors:

```
You are a [AGE_DESC] [GENDER] from
[COUNTRY] [EDUCATION_DESC].
You are politically [ORIENTATION],
[RELIGIOSITY] religious,
and have [TRUST_LEVEL] trust in
institutions.
Answer from this perspective. Respond
with only the number.
```

GSS variant. For GSS, we replace nationality with race, gender, and party affiliation, and use political views and religiosity as attitude anchors.

Question prompts consist of the persona text followed by the item text and a list of numbered options; we randomise the order of options at generation time and restore the original indexing when computing metrics. Following Tzachristas et al. (2025), anchor attitudes are derived from a pre-specified set of calibration attributes/items and mapped to short descriptors. These anchor variables are excluded from the held-out target-item set, and no held-out test response is placed in any persona prompt.

B.2 Response and Item Bases

For ordinal items, the response basis Φ_j encodes each category's normalised position on the response scale:

$$t_{j,k} = \frac{2k - (K_j + 1)}{K_j - 1} \in [-1, 1]. \tag{19}$$

We use a low-dimensional, mean-zero basis capturing directional shifts and extremity/dispersion:

$$\Phi_j[k, 1] = t_{j,k}, \tag{20}$$

$$\Phi_j[k, 2] = t_{j,k}^2 - \frac{1}{K_j} \sum_{k'=1}^{K_j} t_{j,k'}^2. \tag{21}$$

For binary items ($K_j = 2$), the second basis column is identically zero, so the model has one effective degree of freedom while retaining a consistent parameter shape.

To enable transfer to held-out items without using human labels, fixed item embeddings are derived from the base simulator's pooled synthetic attitude structure. We draw a pooled synthetic sample from Q_θ ignoring group labels, compute the pooled synthetic correlation matrix $R_{\text{pool}}^{\text{syn}} \in \mathbb{R}^{|J| \times |J|}$, and take its rank-$r$ eigendecomposition:

$$R_{\text{pool}}^{\text{syn}} \approx U_r \Lambda_r U_r^\top. \tag{22}$$

The embedding of item j is its principal-coordinate vector:

$$e_j = \left(\sqrt{\lambda_1} U_r[j,1], \ldots, \sqrt{\lambda_r} U_r[j,r]\right)^\top. \quad (23)$$

Thus e_j places each item in a low-dimensional latent attitude space induced by the simulator's own correlations. The embeddings are recomputed for each base simulator.

B.3 Metric Implementations

We briefly describe implementation choices for the evaluation metrics in Section 2.5.

Ordinal items are encoded as consecutive integers in the natural order of response labels. For nominal items without an inherent order we compute KL divergence but omit EMD and treat categories as unordered.

Correlation matrices R_g^{real} and R_g^{syn} are computed using Pearson correlation between encoded scores. When items have different scales, we z-score responses at the item level before correlation. In addition to the degree-based SAF score in Equation 5, we also experimented with spectral variants that compare the leading eigenvalues and eigenvectors of R_g^{real} and R_g^{syn}; these yield similar qualitative conclusions.

For the main experiments we set $\alpha = 0.5$ in Equation 5, giving equal weight to Frobenius-norm similarity and degree-correlation similarity. We compute the rescaled $\widetilde{\Delta_g^{\text{F}}} \in [0,1]$ by min–max normalising Δ_g^{F} across all groups and simulators within a dataset:

$$\widetilde{\Delta_g^{\text{F}}} = \frac{\Delta_g^{\text{F}} - \min_{g',m} \Delta_{g',m}^{\text{F}}}{\max_{g',m} \Delta_{g',m}^{\text{F}} - \min_{g',m} \Delta_{g',m}^{\text{F}} + \eta},$$

where the minimum and maximum are taken over all groups g' and model configurations m considered on that dataset, and η is a small constant for numerical stability. This makes SAF values comparable across simulators within the same dataset while leaving cross-dataset comparisons qualitative.

Earth Mover's Distance between $P_{g,j}$ and $\hat{P}_{g,j}$ is computed on the discrete support $\{1, \ldots, K_j\}$ with unit spacing. For between-group distortion we restrict to group pairs (g, g') where both have sufficient sample size to estimate means and correlations reliably; we ignore pairs where all responses to an item take the same value in both groups.

B.4 SAF Component Breakdown

To address concerns about the composite nature of SAF, we report its raw components alongside the aggregate score. Table 6 shows the three structural metrics for the guided persona simulator and FairSynth on WVS7.

Method	$\Delta^{\text{F}}\downarrow$	$\rho^{\text{deg}}\uparrow$	$\rho^{\text{corr}}\uparrow$
Guided persona	2.84±.42	.71±.09	.68±.08
+ hist. oracle	3.52±.61	.58±.14	.54±.12
+ temp. scal.	3.18±.48	.64±.11	.61±.10
+ FairSynth	**2.31**±.35	**.79**±.07	**.76**±.06

Table 6: SAF component breakdown (WVS7, mean±SD across groups). Δ^{F}: raw Frobenius distance (lower is better); ρ^{deg}: degree correlation; ρ^{corr}: entry-wise Spearman correlation of upper-triangular R entries. FairSynth improves all three components.

The entry-wise correlation ρ^{corr} provides the most interpretable measure: it directly answers "how well does the simulator preserve the pattern of item–item associations?" FairSynth improves ρ^{corr} from 0.68 to 0.76, indicating strong monotonic agreement in the pattern of pairwise item associations. The histogram oracle, despite achieving near-zero marginal error, degrades ρ^{corr} to 0.54, confirming that marginal-only optimisation harms structure.

SAF normalisation protocol. To ensure reproducibility, we define the normalisation reference set as the fixed baselines in Table 2. For comparisons with new models, we recommend including the same reference set or reporting raw Δ_g^{F} values directly. We also provide an absolute variant $\text{SAF}_g^{\text{abs}} = 1 - \Delta_g^{\text{F}}/(\sqrt{2}|J|)$, which normalises by the maximum possible Frobenius distance for correlation matrices; this yields similar relative rankings (Spearman $\rho > 0.95$ with min–max SAF) while being reference-set invariant.

B.5 Calibration Optimisation

We optimise the shared group parameters $B = \{B_g\}_{g \in \mathcal{G}}$ in Equation 15 with stochastic gradient descent using automatic differentiation. A generic training loop is:

1. Precompute fixed item embeddings e_j and response bases Φ_j for all items (Section 2.6).
2. Initialise $B_g = 0$ for all groups.
3. Repeat until convergence:

(a) Sample a mini-batch of groups and synthetic respondents.

(b) For each respondent and item, obtain base probabilities $q_{i,j}$ from the simulator, compute logit shifts $a_{g,j} = \Phi_j B_g e_j$, and compute $q'_{i,j}$ via Equation 13.

(c) Update empirical estimates of $\tilde{P}_{g,j}(B)$ and $R_g^{\text{cal}}(B)$ from the mini-batch.

(d) Compute $\mathcal{L}_{\text{marg}}$, $\mathcal{L}_{\text{struct}}$, $\mathcal{L}_{\text{fair}}$, and $\mathcal{L}(B)$.

(e) Take an optimisation step on B and apply a Frobenius norm constraint to keep $\|B_g\|_F$ bounded.

In practice we use a small number of epochs over the synthetic sample and select hyperparameters (learning rate, embedding rank r, basis dimension d, loss weights, norm bound) on held-out respondents. Because the calibration operates only on B, it can be reused across different downstream analyses of the same population. For the GRE-based fairness term we use the standard representation of EMD on a one-dimensional ordered support as the ℓ_1 distance between cumulative distribution functions. Concretely, for an ordinal item j we compute $\text{EMD}(P_{g,j}, \tilde{P}_{g,j}(a))$ as the sum of absolute differences between the cumulative probabilities of $P_{g,j}$ and $\tilde{P}_{g,j}(a)$ across categories. In practice we use a smooth approximation to the absolute value to obtain stable gradients, and we backpropagate through the resulting differentiable approximation.[1]

B.6 Alternative Fairness Objectives

The variance-based fairness term $\mathcal{L}_{\text{fair}}(B) = \text{Var}_g(\text{GRE}'_g(B))$ captures dispersion of representation error across groups. We also experimented with several alternative objectives:

Max-GRE objective. $\mathcal{L}_{\text{fair}}^{\max}(B) = \max_g \text{GRE}'_g(B)$ directly targets worst-group error. In practice we use a soft-max approximation for differentiability.

Weighted variance objective. To address the concern that small groups may have unreliable GRE estimates, we define a sample-size-weighted variant:

$$\mathcal{L}_{\text{fair}}^{(w)}(B) = \sum_g w_g \left(\text{GRE}'_g(B) - \overline{\text{GRE}}\right)^2, \quad (24)$$

where $w_g \propto \min(1, \sqrt{n_g}/c)$ with n_g the number of respondents in group g and c a normalising constant. This down-weights groups with very small sample sizes (where GRE estimates are noisy) while preserving the parity objective.

Minority-weighted objective. We also tested a variant that assigns higher weight to historically marginalised groups (e.g., Global South countries, minoritised racial groups) based on external annotations.

These alternatives can be plugged into Equation 15 without changing the rest of the framework.

Table 7 compares these fairness objectives on WVS7. Using max GRE as the fairness target slightly reduces worst-group error but at the cost of higher average KL; weighting by sample size provides more stable optimisation for small groups without changing conclusions for large groups. We adopt the unweighted variance-based objective as our default because it provides a good balance across metrics and does not require specifying group-specific weights or external annotations.

Objective	KL↓	SAF↑	Std↓	Max↓
Variance (ours)	**.09**±.03	**.82**±.05	.028±.01	.19
Max GRE	.10±.03	.80±.06	**.022**±.01	**.17**
Weighted Var.	.09±.03	.81±.05	.027±.01	.18
Minority-wtd.	.10±.03	.81±.05	.026±.01	.18

Table 7: Fairness objectives (WVS7, mean±SD across 3 seeds). Variance provides best overall balance; weighted variants yield similar results with slightly better stability for small groups.

C Additional Experimental Details

C.1 Group Definitions

For WVS7 we start from the full set of respondents with complete data on the selected items and demographics. Age is partitioned into three bands (18–34, 35–54, 55+), and education is partitioned into coarse levels (primary or less, secondary, tertiary). Groups are defined as country × gender × age × education, with small cells merged into an "other" category per country. This produces 156 merged demographic strata covering all 64 WVS7 countries; rare demographic cells are sampled or merged within country to ensure the minimum group size while preserving country-level stratification.

For the within-country survey we construct intersectional groups such as race × gender × party identification, again applying a minimum-size rule

[1] We found that this approximation closely tracks the exact EMD values used for reporting GRE while avoiding vanishing gradients when group-level distributions are already well aligned.

and merging rarely observed combinations. We found that FairSynth metrics and calibration behaviour were stable across a range of reasonable binning schemes; sensitivity checks with alternative bin boundaries led to qualitatively similar patterns.

C.2 Item Selection and Preprocessing

We include only items that satisfy three criteria: they appear with the same wording and response categories for all populations under consideration; they have at least two distinct response values in each group; and they correspond to attitudes or values rather than purely factual or demographic questions. For each dataset this yields a set J of 19–26 items spanning values, trust and social issues.

Responses coded as "don't know", refusals and missing values are treated as missing and excluded from the estimation of group-level distributions. For ordinal metrics we map response categories to consecutive integers in the natural order induced by the questionnaire; for nominal items we use one-hot encodings and omit EMD-based metrics.

C.3 Base Model Details

In all experiments the base simulator Q_θ is instantiated using large instruction-tuned language models. We evaluate the following models:

Prompt-based simulators. For the zero-shot and persona-based simulators we use open-source models:

- **Qwen2.5-72B-Instruct** served locally on 4×A100 GPUs with vLLM, using $T = 0.3$ and top-$p = 0.95$;
- **Qwen2.5-7B-Instruct** for resource-constrained settings, requiring only 16GB VRAM;
- **Llama-3.1-8B-Instruct** as an alternative 8B model with similar sampling parameters.

All models are queried with these fixed temperature and nucleus sampling parameters, and responses are truncated to a single token representing the option index. Unless otherwise noted, main-text results use Qwen2.5-72B as the base model; we report cross-model comparisons in Table 8 below.

Distributionally specialised simulator. For the distributionally specialised simulator we follow the objective of Cao et al. (2025), fine-tuning Llama-3-8B-Instruct using LoRA adapters (rank 16, $\alpha = 32$) to minimise divergence between predicted and empirical response distributions on a subset of groups and items. Training uses AdamW with learning rate 2×10^{-5} for 3 epochs. Hyperparameters such as learning rate, number of training steps and regularisation strength are tuned on validation groups. The aim of this baseline is not to compete with FairSynth but to represent a strong method that directly optimises marginal fidelity.

Cross-model comparison. Table 8 shows that FairSynth calibration consistently improves all base models across open-source LLMs of different sizes.

Base Model	Variant	KL↓	SAF↑	GRE↓
Qwen2.5-7B	Guided	.20±.06	.72±.08	.18±.05
Qwen2.5-7B	+ FairSynth	.11±.04	.79±.06	.14±.04
Llama-3.1-8B	Guided	.21±.07	.71±.09	.19±.05
Llama-3.1-8B	+ FairSynth	.12±.04	.78±.06	.15±.04
Qwen2.5-72B	Guided	.17±.05	.76±.07	.16±.04
Qwen2.5-72B	+ FairSynth	**.09**±.03	**.82**±.05	**.12**±.03

Table 8: Cross-model comparison with open-source LLMs (WVS7, mean±SD across 3 seeds). FairSynth improves all base models; larger models achieve better results.

C.4 Calibration Procedure

Response basis and item embeddings. For ordinal items we encode response categories by their normalised position on the response scale,

$$t_{j,k} = \frac{2k - (K_j + 1)}{K_j - 1} \in [-1, 1]. \tag{25}$$

The response-basis matrix uses centred linear and quadratic components:

$$\Phi_j[k, 1] = t_{j,k}, \tag{26}$$

$$\Phi_j[k, 2] = t_{j,k}^2 - \frac{1}{K_j} \sum_{k'=1}^{K_j} t_{j,k'}^2. \tag{27}$$

For binary items ($K_j = 2$), the second basis column is identically zero. To compute e_j, we draw a pooled synthetic sample from Q_θ, compute the pooled synthetic correlation matrix $R^{\text{syn}}_{\text{pool}}$, and take its rank-$r$ eigendecomposition $R^{\text{syn}}_{\text{pool}} \approx U_r \Lambda_r U_r^\top$. We set $e_j = (\sqrt{\lambda_1} U_r[j, 1], \ldots, \sqrt{\lambda_r} U_r[j, r])^\top$. The embeddings are recomputed for each base simulator and are fixed during calibration.

We initialise all group parameters B_g to zero and optimise the multi-objective loss in Equation 15

using Adam with a fixed learning rate. Item embeddings e_j and response bases Φ_j are precomputed once and held fixed throughout training. Base probabilities $q_{i,j}$ are cached once from Q_θ for all persona–item pairs before calibration begins; no repeated LLM calls occur during optimisation. At each update we sample a mini-batch of groups and respondents, compute calibrated probabilities $q'_{i,j}$ via Equations 12–13 using cached $q_{i,j}$, update mini-batch estimates of group-level distributions and correlation matrices, and evaluate the loss components.

To stabilise the structural term we maintain exponential moving averages of covariance estimates per group before converting to correlation matrices. We enforce a Frobenius norm constraint $\|B_g\|_F \leq \rho$ for all g after each update, projecting onto the Frobenius ball of radius ρ when necessary. The radius ρ, embedding rank $r \in \{4, 8, 12\}$, basis dimension $d \in \{1, 2\}$, loss weights λ_{struct}, λ_{fair} and λ_{reg}, and the number of optimisation steps are chosen by grid search on the validation groups, optimising a weighted sum of average GRE and the standard deviation of GRE across groups. The shared parameterisation reduces the total number of learned parameters from $O(|\mathcal{G}||J|\bar{K})$ (e.g., $\sim$11,000 for WVS7 under per-(g, j) shifts) to $O(|\mathcal{G}|rd)$ (e.g., $\sim$2,500 with r=8, d=2), substantially reducing overfitting risk for small groups and accelerating convergence.

C.5 Compute and Sampling

FairSynth requires access to the base simulator's probabilities (or log-probabilities) over response options, rather than just sampled labels. This is readily supported by many modern LLM APIs but may incur additional latency or cost compared to pure sampling.

In our WVS7 configuration, we calibrate 156 merged demographic strata (covering all countries but merging rare demographic combinations) with approximately 500 synthetic respondents per group ($\sim$78,000 total respondents). We cache base logprobs $q_{i,j}$ for all persona–item pairs once ($\sim$0.8M calls, $\sim$160M tokens); calibration then optimises over B using cached outputs with no repeated LLM calls per epoch. With the shared parameterisation ($\sim$2,500 parameters), calibration runs for 30 epochs in 6–8 hours wall-clock on Qwen2.5-72B (4×A100 GPUs). For 7B–8B models, requirements are more modest (single A100, 3–5 hours).

We limit the number of synthetic respondents per group to a fixed budget to keep calibration tractable. Preliminary experiments with larger samples suggested diminishing returns beyond this scale. While the overhead is non-negligible, it is modest relative to fine-tuning a large model from scratch. We therefore view FairSynth as best suited for settings where one plans to reuse a calibrated simulator across multiple studies or as a shared resource.

Training vs. evaluation: differentiability. We clarify how FairSynth handles differentiability. **Training-time estimation:** Base logprobs $q_{i,j}$ are cached once from Q_θ for all (persona, item) pairs; calibration then reuses these cached outputs without repeated LLM calls per epoch. To compute $R_g^{\text{cal}}(B)$ differentiably during training, we use *soft expected scores*: $m_{i,j} = \sum_{k=1}^{K_j} k \cdot q'_{i,j}(k)$, and compute correlations over $m_{i,\cdot}$ across respondents. This is fully differentiable via automatic differentiation. **Evaluation-time estimation:** Marginal metrics (KL, EMD, GRE) use averaged calibrated probabilities $\tilde{P}_{g,j}$. Correlation matrices R_g^{syn} and R_g^{cal} are computed from sampled labels for interpretability. We found expectation-based correlations closely track sample-based correlations at our sample sizes ($\sim$500 per group; sensitivity analysis in Appendix F.2).

D Additional Result Tables

D.1 Standard Deviations for Main Tables

Table 9 reports full standard deviations for Tables 2 and 3.

	WVS7 (SD)			GSS (SD)		
Simulator	KL	SAF	GRE	KL	SAF	GRE
Tabular baseline	.08	.09	.06	.07	.08	.05
Zero-shot LLM	.12	.11	.08	.09	.10	.06
Demog. persona	.07	.08	.05	.06	.07	.04
Guided persona	.05	.07	.04	.04	.06	.03
+ hist. oracle	.01	.12	.01	.01	.11	.01
+ temp. scal.	.02	.09	.04	.02	.08	.03
+ FairSynth	.03	.05	.03	.02	.04	.02

Table 9: Standard deviations for Table 2 (across 3 seeds).

D.2 Specialised Simulator Results

Table 10 provides a fuller breakdown for the distributionally specialised simulator and its FairSynth-calibrated variant.

	WVS7				GSS 2022			
Simulator	KL↓	SAF↑	GRE↓	FI≈1	KL↓	SAF↑	GRE↓	FI≈1
Specialised	.10	.71	.13	.69	.09	.73	.12	.75
+ FairSynth	**.07**	**.80**	**.11**	**.92**	**.06**	**.82**	**.09**	**.94**

Table 10: Specialised simulator with/without FairSynth. Calibration improves all metrics.

E Additional Analysis and Ablation Details

E.1 Group-level Breakdown

Figure 2 summarises region-, education-, and race-level FairSynth gains. WVS7 regions with larger initial errors, especially Africa and Latin America, receive larger GRE reductions than Western Europe or North America. In GSS, Black and lower-education groups see 40–50% GRE reductions, while groups already represented with lower error change less.

Figure 3 provides a per-country view of FairSynth's effects on WVS7. For each country we plot the group-level GRE and FI before and after calibration, averaged over demographic groups within that country.

We observe that almost all countries benefit from calibration, with the largest improvements in countries where base GRE and flattening were most severe. Countries with small numbers of groups or low sample sizes exhibit more variance, but the overall trend mirrors the group-level breakdown above.

E.2 Ablations on Loss Components

We ablate components of the FairSynth loss (Equation 15) by comparing four variants: (i) marginal-only ($\lambda_{\text{struct}} = \lambda_{\text{fair}} = 0$), (ii) marginal + fairness (no structural term), (iii) marginal + structural (no fairness term), and (iv) full FairSynth.

Variant (WVS7)	KL↓	SAF↑	Std↓	Max↓
Guided (base)	.17±.05	.76±.07	.054±.02	.35
Marginal only	**.05**±.02	.65±.09	.048±.02	.31
Marg + fair	.07±.02	.63±.10	**.020**±.01	**.18**
Marg + struct	.10±.03	.79±.06	.038±.01	.28
Full FairSynth	.09±.03	**.82**±.05	.028±.01	.19

Table 11: Ablation on loss components (WVS7, mean±SD across 3 seeds). Bold marks per-column scalar bests; Full objective provides the best overall balance.

Table 12 reports the same analysis for GSS. The pattern closely matches the cross-national results: marginal-only calibration achieves the lowest KL but harms structural attitude fidelity and fairness, while full FairSynth provides a balanced trade-off.

Variant (GSS)	KL↓	SAF↑	Std↓	FI≈1
No calibration	.15±.04	.78±.06	.048±.02	.83
Marginal only	**.04**±.01	.68±.08	.065±.02	1.03
Marg + fair	.06±.02	.71±.07	**.018**±.01	1.00
Marg + struct	.09±.03	**.86**±.04	.042±.02	.85
Full FairSynth	.08±.02	.84±.04	.025±.01	**.96**

Table 12: Ablation (GSS, mean±SD across 3 seeds). Full objective balances all metrics.

E.3 Item-level Generalisation

We partition each questionnaire into a calibration subset J_{calib} (18 items for WVS7, 13 items for GSS) and a held-out subset J_{held} (8 items for WVS7, 6 items for GSS), train shared group parameters B_g using only J_{calib}, and evaluate all metrics separately on both subsets. Because B_g applies to all items via fixed embeddings e_j, held-out items can improve even though their human marginals are never used during training.

Data	**Items**	KL↓	SAF↑	GRE↓
WVS7	J_{calib} (base)	.17±.05	.76±.07	.16±.04
WVS7	J_{calib} (FS)	**.09**±.03	**.83**±.05	**.12**±.03
WVS7	J_{held} (base)	.18±.06	.74±.08	.17±.05
WVS7	J_{held} (FS)	.11±.04	.79±.06	.13±.03
GSS	J_{calib} (base)	.15±.04	.78±.06	.14±.03
GSS	J_{calib} (FS)	**.07**±.02	**.85**±.04	**.10**±.02
GSS	J_{held} (base)	.16±.05	.76±.07	.15±.04
GSS	J_{held} (FS)	.09±.03	.80±.05	.12±.03

Table 13: Item-level generalisation (mean±SD across 3 seeds). FairSynth (FS) improves both calibration and held-out items.

On WVS7, FairSynth reduces average KL divergence on held-out items from 0.18 to 0.11 (a 39% reduction, compared to 47% on calibration items) and reduces mean GRE from 0.17 to 0.13. SAF increases on held-out items from 0.74 to 0.79, although gains are smaller than on calibration items (0.76 to 0.83). GSS shows a similar pattern: FairSynth improves marginal fidelity on J_{held}

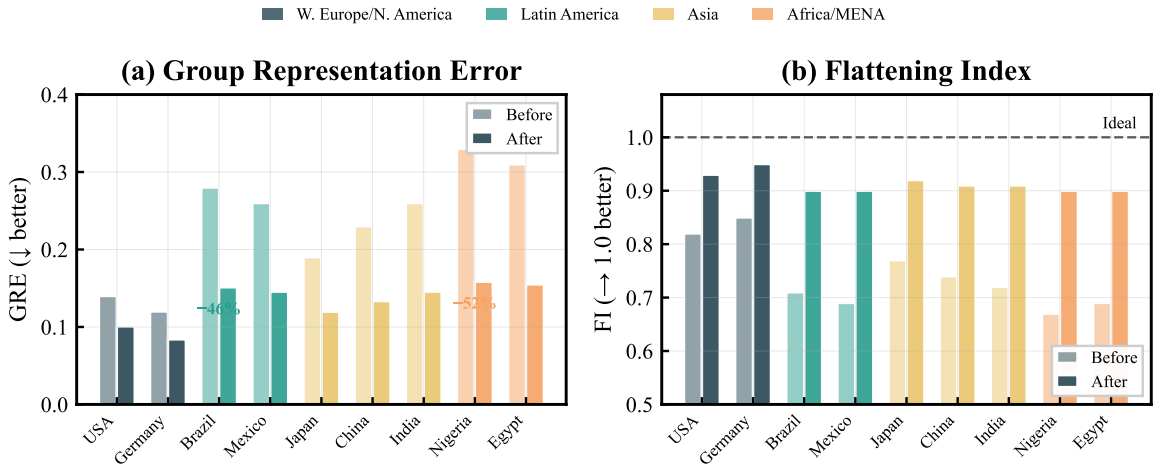


Figure 3: Per-country GRE and FI changes (WVS7). (a) GRE: largest reductions for Nigeria/Brazil (40–50%). (b) FI: moves toward 1.0 for all countries.

(KL from 0.16 to 0.09) while maintaining SAF. All results are averaged over three random seeds; standard deviations are reported in the tables.

E.4 Structural Loss Anchoring Ablation

Table 14 compares two variants of the structural loss term: anchoring to the pre-calibration synthetic correlations R_g^{syn} (our default) versus anchoring directly to the human correlations R_g^{real}.

Anchor	KL↓	SAF↑	Std↓	Max↓
R_g^{syn} (ours)	.09±.03	.82±.05	**.028**±.01	**.19**
R_g^{real}	**.08**±.02	**.84**±.04	.045±.02	.27
Mixed (0.5/0.5)	.08±.02	.83±.04	.036±.01	.23

Table 14: Structural anchoring (WVS7, mean±SD across 3 seeds). R_g^{syn} is more robust for fairness.

Anchoring to R_g^{real} produces marginally higher SAF on average, but the increase in Std(GRE) and max GRE indicates that the optimiser occasionally distorts marginals to match noisy empirical covariances, particularly for small groups. A mixed anchoring strategy (convex combination of the two targets) lies between these extremes. We adopt R_g^{syn} as the default because it provides robust improvements across groups of varying size.

E.5 Hyperparameter Sensitivity Analysis

We analyse the sensitivity of FairSynth to the loss weights λ_{struct} and λ_{fair} by sweeping each over the range $\{0, 0.1, 0.5, 1.0, 2.0, 5.0\}$ while holding the other at its default value ($\lambda_{\text{struct}} = 1.0$, $\lambda_{\text{fair}} = 1.0$). Table 15 reports results on WVS7 with the guided persona base simulator.

λ_s	λ_f	KL↓	SAF↑	Std↓	Max↓
0	1.0	**.06**±.02	.63±.10	**.020**±.01	.18
0.1	1.0	.07±.02	.72±.08	.022±.01	**.17**
0.5	1.0	.08±.03	.79±.06	.025±.01	.18
1.0	1.0	.09±.03	.82±.05	.028±.01	.19
2.0	1.0	.10±.03	.83±.05	.030±.01	.20
5.0	1.0	.12±.04	**.84**±.04	.038±.01	.24
1.0	0	.10±.03	.79±.06	.042±.02	.28
1.0	0.1	.09±.03	.80±.06	.038±.01	.25
1.0	0.5	.09±.03	.81±.05	.032±.01	.22
1.0	1.0	.09±.03	.82±.05	.028±.01	.19
1.0	2.0	.09±.03	.81±.05	**.020**±.01	.17
1.0	5.0	.10±.03	.78±.06	**.020**±.01	**.16**

Table 15: Hyperparameter sensitivity (WVS7, mean±SD across 3 seeds). λ_s/λ_f = structural/fairness weights. Default (1.0, 1.0) provides balanced trade-off.

Key observations. (i) Setting $\lambda_{\text{struct}} = 0$ leads to substantial SAF degradation (0.63 vs. 0.82), confirming that the structural term is essential. (ii) Increasing λ_{struct} beyond 1.0 yields diminishing returns in SAF while increasing KL. (iii) Setting $\lambda_{\text{fair}} = 0$ leaves worst-group GRE high (0.28); moderate values (0.5–2.0) provide the best balance. (iv) Very high λ_{fair} (5.0) aggressively equalises GRE but begins to harm SAF, likely because the optimiser distorts item distributions to force equal error across groups.

We also verified that the regularisation weight $\lambda_{\text{reg}} \in \{0.01, 0.1, 1.0\}$ has minimal effect on final metrics as long as it prevents divergence; we use $\lambda_{\text{reg}} = 0.1$ throughout.

E.6 Structural Attitude Graphs

We visualise item–item correlation networks to qualitatively assess how different simulators pre-

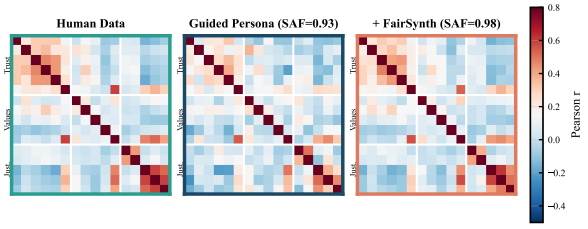


Figure 4: Item–item correlation matrices (US, n=2,596). **Left**: Human data. **Middle**: Guided persona attenuates cross-cluster links. **Right**: FairSynth recovers structure.

serve attitude structure. Figure 5 shows item graphs for three countries spanning different cultural contexts (United States, Germany, Japan), comparing human data with the guided persona simulator and FairSynth.

Figure 6 provides an extended comparison including the histogram oracle and the distributionally specialised simulator for the United States.

F Additional Robustness Analyses

F.1 Survey Weights Sensitivity

Our main experiments treat the realised survey sample as the population of interest and use unweighted responses. Here we test robustness to design weights provided by WVS7 and GSS.

We recompute human group marginals using design weights:

$$P^{(w)}_{g,j}(k) = \frac{\sum_{i \in D_g} w_i \cdot \mathbf{1}[y_{i,j} = k]}{\sum_{i \in D_g} w_i}, \qquad (28)$$

where w_i is the survey weight for respondent i. We then recompute all evaluation metrics (KL, GRE, SAF) with respect to $P^{(w)}_{g,j}$ rather than $P_{g,j}$.

Table 16 compares key metrics under unweighted and weighted evaluation for the guided persona simulator and FairSynth. The direction and magnitude of FairSynth improvements are preserved under weighted evaluation: KL reduction remains 45–50%, GRE reduction 22–28%, and SAF improvements 0.05–0.07. Weighted marginals shift absolute metric values slightly (typically <5% relative change), but the ranking of methods and the relative gains from calibration are stable.

Weighting	Method	KL↓	SAF↑	GRE↓
Unweighted	Guided	.17±.05	.76±.07	.16±.04
	+ FairSynth	.09±.03	.82±.05	.12±.03
Weighted	Guided	.18±.05	.75±.07	.17±.04
	+ FairSynth	.10±.03	.81±.05	.13±.03

Table 16: Survey weights sensitivity (WVS7, mean±SD across 3 seeds). FairSynth improvements are robust to weighted vs. unweighted evaluation.

We also experimented with calibrating FairSynth using weighted human marginals $P^{(w)}_{g,j}$ in the training loss. Results were qualitatively similar, with slightly larger improvements for groups whose weights differ substantially from uniform. Incorporating design weights without overfitting small cells with extreme weights remains an open problem; we leave adaptive weighting schemes to future work.

F.2 Sample Size Sensitivity

We analyse how the number of synthetic respondents per group affects metric stability. For each group g, we vary the synthetic sample size $|\hat{D}_g| \in \{50, 100, 200, 500, 1000\}$ and measure the coefficient of variation (CV) of GRE and SAF across 10 independent samples.

$\lvert\hat{D}_g\rvert$	50	100	200	500	1000
CV(GRE)	.12	.08	.05	.03	.02
CV(SAF)	.15	.10	.06	.04	.03

Table 17: Coefficient of variation of metrics vs. synthetic sample size per group (WVS7, averaged over 20 representative groups).

Metrics stabilise rapidly: CV drops below 0.05 for sample sizes ≥ 200. We use $|\hat{D}_g| = 500$ in all main experiments to ensure reliable correlation estimation while keeping compute tractable.

F.3 SAF Robustness to Correlation Estimator

Our main SAF analysis uses Pearson correlations on integer-coded ordinal responses. To address concerns that this choice may affect conclusions, we recompute SAF components using Spearman rank correlations, which are more robust to ordinal data properties.

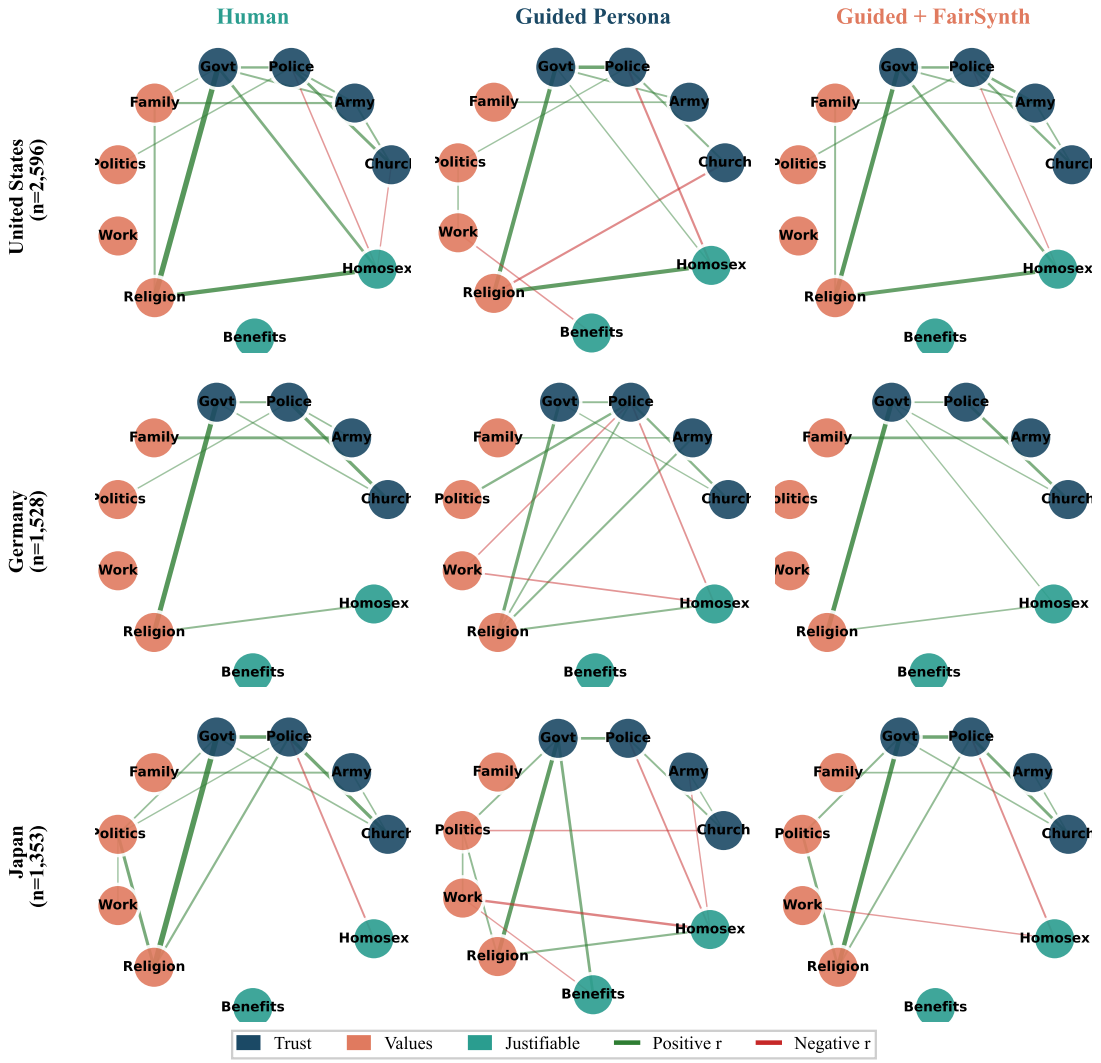


Figure 5: Correlation networks for US/Germany/Japan. Guided persona attenuates cross-cluster edges; FairSynth restores them (SAF +0.06–0.11).

Corr. type	Method	$\Delta^{F}\downarrow$	$\rho^{deg}\uparrow$	$\rho^{corr}\uparrow$
Pearson	Guided	2.84±.42	.71±.09	.68±.08
	+ FairSynth	2.31±.35	.79±.07	.76±.06
Spearman	Guided	2.91±.45	.69±.10	.66±.09
	+ FairSynth	2.38±.38	.77±.08	.74±.07

Table 18: SAF components under Pearson vs. Spearman correlations (WVS7). Relative improvements from FairSynth are consistent across both estimators.

The relative improvement from Guided to FairSynth is preserved under Spearman correlations: Δ^{F} decreases by 18% (vs. 19% for Pearson), ρ^{deg} increases by 0.08 (vs. 0.08), and ρ^{corr} increases by 0.08 (vs. 0.08). We also verified that polychoric correlations (computed via maximum likelihood for a subset of groups with sufficient cell counts) yield similar relative rankings, though absolute values differ slightly due to the latent-variable assumption underlying polychoric estimation.

F.4 Sign-Flip Threshold Sensitivity

Sign-flip rates can be inflated by group-pair–item triples where $|\Delta^{real}_{g,g',j}|$ is very small. Table 19 reports flip rates under different dead-zone thresholds δ, excluding triples where $|\Delta^{real}| < \delta$.

Method	δ=0	δ=0.05	δ=0.10	δ=0.15
Guided persona	.11	.09	.07	.05
+ hist. oracle	.20	.17	.14	.11
+ FairSynth	**.05**	**.04**	**.03**	**.02**

Table 19: Sign-flip rate (WVS7) under different dead-zone thresholds. FairSynth maintains lower flip rates across all thresholds; the relative reduction vs. Guided (55%→60%) is stable.

At δ=0.05, approximately 12% of triples are excluded. The relative improvement from FairSynth

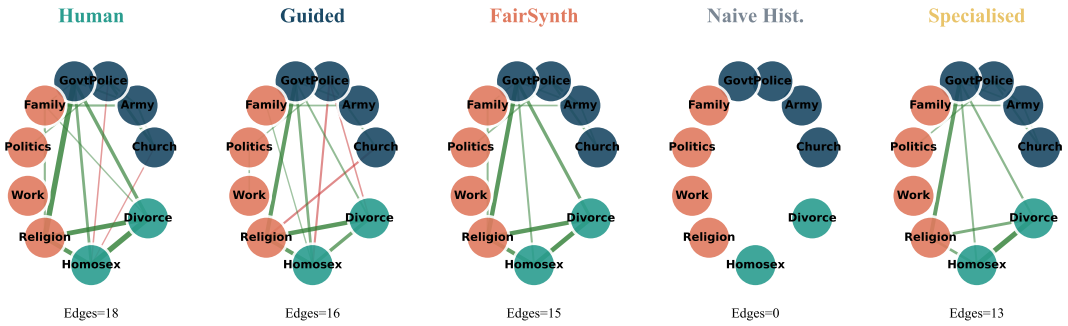


Figure 6: Extended comparison (US). Histogram oracle distorts structure; specialised simulator partially flattens clusters; FairSynth balances marginals and structure.

is consistent across thresholds, confirming that the reduction in sign flips is not driven by near-zero differences. The histogram-oracle row should be read as a finite-sample diagnostic for independent-marginal sampling, not as a deployable simulator.

F.5 Answer-only Generation and Option Probability Extraction

All experiments use an answer-only multiple-choice protocol. Prompts end with an instruction such as *"Respond with only the number"* or *"Respond with only the option label"*; generation is limited to a single token. We then extract option probabilities $q_{i,j}(k)$ from the LLM's first-token log-probabilities over the valid labels, rather than from unconstrained free-form text. Calibration acts only on this renormalised answer distribution.

For each item with K_j options, we present numbered options (1, 2, …, K_j) and extract log-probabilities for the corresponding digit tokens. To handle tokenisation variation:

- we query log-probabilities for both "1" and " 1" (with leading space) and take the higher probability;
- for $K_j \geq 10$, we use letter labels (A, B, …) to avoid multi-token labels;
- we renormalise the extracted probabilities over the valid labels to form a distribution over $\mathcal{Y}_j$.

All WVS7 items have $K_j \leq 10$; for any item with ten or more categories, including two GSS items with 11+ categories, we use letter labels. Because the output contains no rationale text, the calibrator cannot create a mismatch between a written explanation and a final answer. This restriction is central to the scope of FairSynth.

Check	Purpose and implementation
Valid-label coverage	Verify that every admissible option label has an extracted log-probability for each persona–item pair; rerun with a larger log-probability budget or letter labels when coverage fails.
Valid-label mass	Record the pre-renormalisation probability mass assigned to valid option labels. Low mass flags prompts for which the model prefers invalid continuations despite the answer-only instruction.
Tokenisation aliases	Compare aliases such as "1" vs. " 1" and numeric vs. letter labels on a small audit subset. Large shifts indicate tokenizer-specific artefacts rather than survey-response differences.
Sampling agreement	For a small random audit set, compare the renormalised option probabilities with empirical frequencies from repeated constrained one-token sampling; report JS/KL divergence or rank correlation.
Label and order robustness	Re-run a held-out subset with alternative numeric/letter labels and option-order permutations, then compare group marginals and GRE after mapping labels back to the original scale.
Free-form leakage	Verify that sampled completions contain only one valid option label and no rationale, prose prefix, or multi-turn continuation. Leakage indicates that the example falls outside FairSynth's answer-only scope.

Table 20: Diagnostic protocol for the answer-only probability interface. These checks do not change the FairSynth objective; they document whether first-token label probabilities are an internally consistent categorical-option interface for a given model and prompt template, without validating unconstrained free-form answer generation.

Diagnostic checks for first-token probabilities. Because first-token probabilities can be unreliable outside an explicit answer-only protocol (Wang et al., 2024), FairSynth treats them only as an interface for constrained option-choice simulation. To make this interface auditable, new deployments

should report the following diagnostics. They are designed to detect tokenisation, label-set, and option-order failures before calibration is interpreted.

Passing these diagnostics supports only the restricted answer-only setting studied in this paper. It should not be interpreted as evidence that first-token probabilities align with answers in unconstrained rationale-first or free-form survey protocols.

F.6 Group Construction and Merging Protocol

For WVS7, we begin with candidate cells defined by country $\times$ gender $\times$ age $\times$ education. We apply a minimum group size threshold of $n_{\min} = 30$ respondents and merge cells below this threshold into an "other" category within each country, preserving country-level stratification. The final evaluation uses 156 merged demographic strata covering all 64 countries. Table 21 shows the distribution of final group sizes.

Statistic	Min	Median	Mean	Max
WVS7 group size	32	412	567	2,841
GSS group size	48	186	202	892

Table 21: Distribution of group sizes after merging.

For GSS, groups are race $\times$ gender $\times$ party, with the same $n_{\min} = 30$ threshold. Rare combinations (e.g., "Other race $\times$ Independent") are merged into adjacent categories based on demographic similarity.

F.7 Nominal vs. Ordinal Item Handling

Of the 26 WVS7 items and 19 GSS items used in our experiments, 24/26 and 17/19 respectively are ordinal (Likert-type or ordered categorical). For the 2 WVS7 nominal items (categorical without inherent order), we:

- Compute KL divergence (order-invariant) but omit EMD from GRE;
- Exclude these items from correlation-based SAF computation.

Sensitivity analysis excluding all nominal items yields nearly identical results (mean KL changes by <0.01), confirming that our conclusions are driven by ordinal items.