Country Name or Cultural Evidence? A Causal Audit of Population-Grounded Steering in Large Language Models

Ziyan Yang
Jinan University
KerryYang@stu2022.jnu.edu.cn

Kuo Zhao
Jinan University
zhaokuo@jnu.edu.cn

Abstract

When a country name and an observed population distribution point in different directions, which does a large language model follow? We introduce a four-condition counterfactual audit comprising a no-cue baseline, a country label, matched population evidence, and a conflict condition in which the label and evidence refer to different populations. On a quality-controlled diagnostic subset of GlobalOpinionQA, we select 100 population pairs and evaluate both reciprocal directions, yielding 200 directed contrasts and 3,200 likelihood-based predictions from four instruction-tuned models in the 7B–9B range. Country labels yield modest mean gains in fit to the named population (Country Influence: 0.028–0.071), whereas direct population evidence yields substantially larger gains (Evidence Influence: 0.184–0.284). The paired evidence advantage is positive for every model. Under conflict, normalized Evidence Override ranges from 0.543 to 0.728, 85.5–97.5% of predictions are closer to the evidence population, and all four models agree on the evidence side for 161 of 200 contrasts. These results establish a clear hierarchy of steering signals: country names influence model behavior, but explicit population distributions exert far greater influence and usually prevail when the two disagree. The pattern is consistent with evidence-weighted updating from a weaker, defeasible country cue. Our audit provides an evaluation protocol for pluralistic alignment, connecting steerable and distributional pluralism by testing not merely whether a model can be steered toward different populations, but whether its behavior tracks the observed plurality within those populations.

1 Introduction

Suppose a large language model (LLM) is asked to predict how respondents from Indonesia would answer an opinion survey question, and its predicted distribution changes. What does that shift mean? The model may be drawing on population-relevant regularities, or it may simply be reacting to the word *Indonesia*. Standard country-conditioned evaluations cannot tell these explanations apart.

Our audit provides an evaluation protocol for pluralistic alignment, connecting steerable and distributional pluralism by testing not merely whether a model can be steered toward different populations, but whether its behavior tracks the observed plurality within those populations. A useful distinction is between *steerable pluralism*, in which a model is directed toward a specified perspective, and *distributional pluralism*, in which its behavior is calibrated to the distribution of views within a target population (Sorensen et al., 2024). Country-conditioned prompting appears to connect the two: a country name specifies a perspective, and the resulting prediction can be compared with survey responses from that country. Prior work shows that country, language and demographic cues can shift LLM responses toward some human populations (DURMUS et al., 2024; AlKhamissi et al., 2024). Yet a country name may activate useful population-level knowledge, a generic persona, geopolitical salience, linguistic associations, or a coarse national prior. A shift under the label therefore establishes label responsiveness, instead of population grounding.

Direct population evidence makes this ambiguity testable. A survey distribution is an imperfect, sample-dependent aggregate, but unlike a bare label it records how respondents answered the specific question. This comparison separates three capabilities that are often bundled together as cultural steerability. *Label responsiveness* asks whether naming a country changes the prediction. *Evidence compliance* asks whether an observed response distribution moves the prediction toward that population. *Cue arbitration* asks which signal governs behavior when the label and the evidence point to different populations. The conflict condition pro-

Proceedings of the First Workshop on Pluralistic Value Alignment of LLMs, pages 176–186
November 10, 2026 ©2026 Association for Computational Linguistics

vides the decisive test: when the name and the numbers disagree, which does the model follow?

We introduce a four-condition counterfactual audit that compares a no-cue baseline, a country label, matched population evidence, and a label–evidence conflict while holding fixed the survey item, semantic options, displayed option order, model and scoring rule. The conflict condition turns an otherwise ambiguous shift into a cue-arbitration test.

On a quality-controlled, high-divergence subset of GlobalOpinionQA, we evaluate 100 population pairs in both directions across four instruction-tuned open-weight models, yielding 200 directed contrasts and 3,200 likelihood-based predictions. Stable option-label likelihoods provide comparable response distributions without free-text generation or post-hoc parsing.

Across all four models, country labels yield modest mean improvements in population fit, whereas direct population evidence produces much larger gains. When the cues conflict, 85.5–97.5% of predictions are closer to the evidence population, and all four models agree on the evidence side for 161 of 200 contrasts. This ordering is shared across models, although the strength of updating varies by model and item. Country labels are therefore informative cues, but not substitutes for population evidence.

Our contributions are summarized as follows.

- **A diagnostic formulation of population grounding.** We distinguish label responsiveness, evidence compliance and cue arbitration, showing why country-conditioned change alone cannot establish that a model tracks the observed distribution of views within a population.

- **A controlled counterfactual evaluation protocol.** We introduce a four-condition reciprocal audit for pluralistic alignment that isolates country and evidence influence within the same model and directly tests their competition under conflict. We operationalize the design using three likelihood-based measures: Country Influence, Evidence Influence and Evidence Override.

- **A cross-model hierarchy of steering signals.** Across four open-weight model families, direct population evidence exerts substantially greater influence than country labels and usually prevails when the cues disagree. The shared direction, together with systematic model- and item-level variation, connects steerable pluralism to the stronger requirement of distributional pluralism.

2 Related Work

We review two intersecting literatures. Work on pluralistic alignment and population simulation asks whose values language models represent and how they can be steered toward a target population. Work on survey elicitation asks whether observed alignment reflects that target or the measurement procedure. Their intersection motivates our central distinction: label responsiveness is not population grounding.

2.1 Pluralistic Value Alignment and Population Simulation

Pluralistic alignment broadens the target of alignment from a single aggregate preference function to multiple reasonable values, groups and distributions (Sorensen et al., 2024). Frameworks such as Modular Pluralism operationalize Overton, steerable, and distributional pluralism through community-specialized models (Feng et al., 2024). Survey benchmarks operationalize the distributional dimension by comparing model predictions with observed response distributions. GlobalOpinionQA introduced a large cross-national collection based on Pew Global Attitudes and World Values Survey items and showed that LLMs represent some populations more closely than others (DURMUS et al., 2024). Related work has examined whose opinions models reflect by default (Santurkar et al., 2023), whether demographic conditioning can reproduce subgroup responses (Argyle et al., 2023), how language, pretraining mixtures, demographic prompts and cultural identity cues affect population alignment (AlKhamissi et al., 2024; Tao et al., 2024), and how finer demographic constraints can model within-country heterogeneity (Zhu et al., 2026). These studies show that population cues can improve fit, but cue sensitivity and average fit do not guarantee faithful representation. Other evaluations find unstable fidelity, compressed within-group variation, and risks of misportraying or essentializing simulated populations (Bisbee et al., 2024; Wang et al., 2025).

Population information can also enter through adaptation or external evidence. Survey-derived supervision can specialize models to reproduce

country- and subgroup-level response distributions (Cao et al., 2025; Suh et al., 2025), while narrative augmentation tests whether survey-based adaptation transfers beyond the original questionnaire format (Adilazuarda et al., 2025). Among evidence-based alignment methods, Evi-DA retrieves World Values Survey distributions to improve country-conditioned prediction under cultural and domain shift (Pham et al., 2026). These methods optimize population fidelity. Our audit instead holds model parameters fixed and identifies which inference-time signal drives prediction. Complementary work shows that matching item-level marginals may still fail to preserve the cross-item value structure observed in human populations (Williams et al., 2026).

The closest concurrent work randomizes survey-country metadata and finds that even randomly assigned labels can redirect individual-response forecasts, whereas verified country metadata can improve accuracy (Lyu et al., 2026). We instead pair a country label with a conflicting observed distribution for the same question. To our knowledge, prior work has not systematically examined label–evidence competition using matched reciprocal contrasts designed to disentangle label responsiveness, evidence compliance and cue arbitration.

2.2 Survey Elicitation and Evaluation Validity

Survey-style evaluation supplies human reference distributions, but its conclusions depend on how responses are elicited. LLM survey behavior can vary with prompt wording, answer order, role instructions and response format (Tjuatja et al., 2024; Dominguez-Olmedo et al., 2024). Constrained and open-ended elicitation can yield different value profiles, while option reordering and paraphrasing can change conclusions about cultural alignment (Röttger et al., 2024; Kabir et al., 2025). Apparent cultural representation may therefore partly reflect stochastic or presentation-dependent behavior (Khan et al., 2025; Libovický, 2026).

Likelihood-based elicitation raises a related measurement issue. First-token probabilities can diverge substantially from the answers expressed in generated text (Wang et al., 2024). We therefore score complete option-label sequences under a fixed within-contrast permutation rather than relying on first-token approximations or free-text parsing. Section 3.4 describes the elicitation procedure in detail.

3 Counterfactual Audit Design

We implemented the distinction between country labels and population evidence through four controlled prompt conditions. The survey question, semantic options, displayed option order, model, and scoring rule remain fixed, while the supplied population cue changes.

3.1 Problem Formulation and Directed Contrasts

We used *country label* for the country or territory name inserted into a prompt, *population evidence* for an observed survey response distribution supplied to the model, and *human reference distribution* for the corresponding normalized survey aggregate used as an evaluation target. The distinction between population evidence and a human reference concerns their roles in the experimental design, not differences in the underlying survey data. The same survey aggregate functions as evidence when shown in the prompt and as a reference when withheld and used only for evaluation.

We used *population fit* to describe proximity to a human reference distribution. We defined *population-grounded steering* as changes in model behavior that tracked the corresponding human reference distribution, thereby distinguishing grounding from simple label responsiveness. In this audit, population grounding was assessed as a graded behavioral property through movement toward the matched reference and relative proximity under conflict. These terms deliberately avoided treating an aggregate survey sample as a complete or normative account of culture.

For question q, let p_L and p_E denote the human reference distributions for the label population L and the evidence population E. An experimental unit is the directed contrast

$$u = (q, L, E), \qquad L \neq E.$$

Direction matters because the label and evidence roles are not interchangeable. The interventions (q, L, E) and (q, E, L) use the same question and population pair but reverse which population is named and which supplies the conflicting evidence. Evaluating both reciprocal directions prevents either population from appearing only as the label or only as the evidence source, and it preserves directional asymmetries in cue weighting. We retain the complete directed key throughout inference, check-

pointing, and analysis so that reciprocal directions are never collapsed.

3.2 Data and Contrast Construction

GlobalOpinionQA contains 2,556 questions drawn from the Pew Research Center's Global Attitudes surveys and World Values Survey Wave 7 (DURMUS et al., 2024). We began with the 1,527 questions in our project export that were classified as subjective or value-relevant. We retained only country or territory samples that could be mapped to an exact shared schema of substantive answer options. We excluded non-national and superseded national samples, malformed distributions, unresolved country-specific placeholders, relative-time questions lacking sufficient survey-wave context, and samples with incompatible option mappings.

Non-substantive responses such as "don't know" and "refused" were removed before renormalization. Substantive responses such as "both", "neither", or "depends" remained when they were part of the original survey scale. Quality control left 1,038 questions and 18,801 question–population distributions.

For each question, we computed the base-2 Jensen–Shannon divergence between every eligible unordered pair of population distributions. We retained at most the three most divergent pairs per question and then selected the 100 most divergent population pairs overall. Selecting high-divergence pairs increased diagnostic resolution: when p_L and p_E were nearly identical, label-following and evidence-following predictions were difficult to distinguish. The per-question cap limited overrepresentation by questions with many eligible population samples.

Evaluating each pair in both directions yielded 200 directed contrasts spanning 44 questions and 55 national or territorial sample labels. Of the 100 unordered pairs, 73 came from Global Attitudes and 27 from WVS. The resulting high-divergence subset was designed to maximize diagnostic resolution for cue arbitration.

3.3 Four Cue Conditions

For each directed contrast (q, L, E), we evaluated the four conditions in Table 1. The baseline contained neither a country label nor population evidence. The country-label condition named L. The population-evidence condition supplied p_L as percentages while withholding the country name. The conflict condition named L while supplying the distribution p_E from a different population.

Condition	Country cue	Survey cue	Diagnostic role
Baseline	–	–	Default prediction
Country label	L	–	Label responsiveness
Population evidence	–	p_L	Evidence compliance
Conflict	L	p_E	Cue arbitration

Table 1: The four cue conditions for a directed contrast (q, L, E).

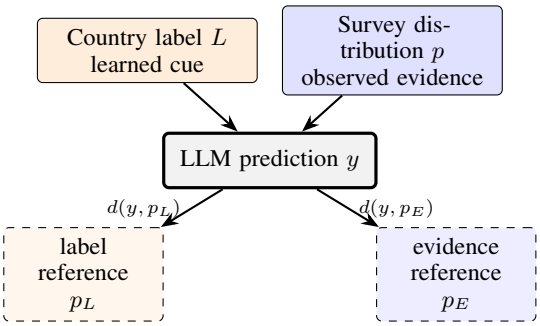


Figure 1: Country labels and observed population distributions are distinct steering signals. Under conflict, relative distance to p_L and p_E indicates which cue the prediction follows more closely.

The design addresses an identification problem. When a country label and population evidence agree, movement toward their shared target is compatible with either label responsiveness or evidence use. In the conflict condition, the two cues imply different targets, so relative distance to p_L and p_E reveals which signal has greater behavioral influence. Because every directed contrast was evaluated under all four conditions, cue effects were estimated through within-contrast comparisons while the underlying item remained fixed. The central test was therefore not merely whether a model could read percentages, but whether explicit evidence displaces a country cue whose influence was measured separately in the matched label condition.

Figure 1 provides a schematic overview of the competing cues and the reference distances used to quantify their relative influence.

Because the joint-cue condition was intentionally incongruent, the design was a controlled four-condition audit centered on cue arbitration. We used *causal* in a limited interventional sense: within the tested model–prompt system, the prompt cue was manipulated while the underlying survey item remained fixed, so differences across condi-

tions identified the cue's effect on model outputs.

The task asked the model to predict which option a randomly sampled respondent would select, given the information supplied in the corresponding condition. Survey percentages are associated with the displayed option labels after permutation. The same displayed order is used for all four conditions of a directed contrast and for every model, so option order cannot covary with either condition or model.

3.4 Models and Likelihood-Based Elicitation

We evaluated Qwen2.5-7B-Instruct (Qwen Team, 2025), Llama-3.1-8B-Instruct (Grattafiori et al., 2024), Mistral-7B-Instruct-v0.3 (Jiang et al., 2023), and Gemma-2-9B-IT (Gemma Team, 2024). All runs used checksum-pinned Q4_K_M GGUF checkpoints and the same pinned `llama.cpp` revision. All models were evaluated with a shared inference backend to standardize runtime conditions, and each model's GGUF-embedded instruction template was applied and independently validated to preserve model-specific chat formatting. Exact repositories, revisions, checkpoint hashes, template hashes, and runtime records were provided in the supplementary artifacts.

Each semantic option was displayed under a stable label $A, B, C, \ldots$. We did not generate or subsequently parse free-form answers. For prompt x and candidate label sequence $a = (a_1, \ldots, a_{|a|})$, we computed its conditional sequence log-likelihood at the common assistant answer position:

$$s(a) = \sum_{j=1}^{|a|} \log p(a_j \mid x, a_{<j}),$$

$$\hat{p}(a \mid x) = \frac{\exp s(a)}{\sum_{a'} \exp s(a')}.$$

Scoring the complete sequence is necessary when a displayed label is represented by multiple tokens. The normalized label probabilities were then mapped back to the original semantic answer options. The 200 directed contrasts and four conditions produced 800 predictions per model and 3,200 predictions overall.

3.5 Measures and Statistical Analysis

Let y_0, y_L, y_P, and y_C denote the model distributions under the baseline, country-label, population-evidence, and conflict conditions. We define base-2 Jensen–Shannon divergence and its square-root distance as

$$D(p, q) = \mathrm{JSD}_2(p, q), \qquad d(p, q) = \sqrt{D(p, q)}.$$

We use Jensen–Shannon divergence because it is symmetric, bounded, and finite when an option receives zero probability. Its square root is a metric.

Country Influence (CI) and Evidence Influence (EI) measure reductions in divergence from the same target distribution p_L:

$$\mathrm{CI} = D(y_0, p_L) - D(y_L, p_L), \tag{1}$$

$$\mathrm{EI} = D(y_0, p_L) - D(y_P, p_L). \tag{2}$$

Positive values indicate that the corresponding cue moves the prediction closer to the target population. We additionally report the paired evidence advantage

$$\mathrm{EI} - \mathrm{CI} = D(y_L, p_L) - D(y_P, p_L),$$

which directly compares the two cue-induced predictions for the same target unit because CI and EI share the same baseline and target.

For the conflict condition, the unnormalized signed contrast is

$$\mathrm{EO}_{\mathrm{raw}} = D(y_C, p_L) - D(y_C, p_E).$$

Positive values indicate an evidence-side prediction, whereas negative values indicate a label-side prediction. Because the magnitude of this raw contrast depends on the separation between p_L and p_E, we also report normalized Evidence Override:

$$\mathrm{EO}_{\mathrm{norm}} = \frac{d(y_C, p_L) - d(y_C, p_E)}{d(p_L, p_E)}.$$

It equals $+1$ when $y_C = p_E$, -1 when $y_C = p_L$, and 0 when the conflict prediction is equidistant from the two reference distributions. Because d is a metric, the reverse triangle inequality guarantees $\mathrm{EO}_{\mathrm{norm}} \in [-1, 1]$, making the measure comparable across population pairs with different separation.

A target unit (q, L) can occur in more than one directed contrast with different evidence populations. We therefore compute CI, EI, and EI $-$ CI once over the 144 unique target units. Evidence Override depends on both L and E and is computed over all 200 directed contrasts.

We estimate 95% percentile intervals using 10,000 bootstrap replicates clustered by

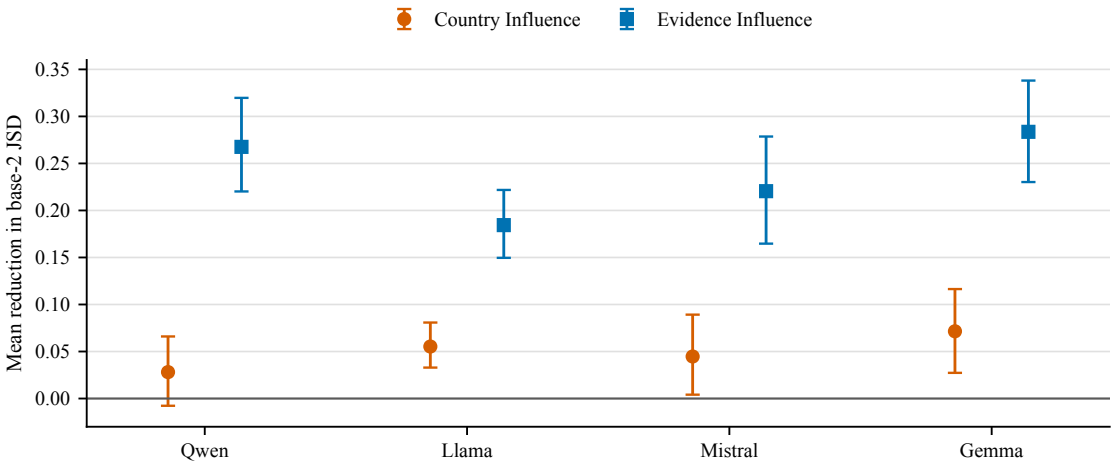


Figure 2: Country-label and population-evidence effects on fit to the target distribution p_L. Points show mean reductions in base-2 Jensen–Shannon divergence relative to baseline. Bars show 95% question-clustered bootstrap intervals. Higher values indicate greater improvement in population fit.

Model	Country Influence	Evidence Influence	Evidence advantage (EI − CI)
Qwen2.5-7B	0.0281 [−0.0077, 0.0660]	0.2676 [0.2202, 0.3197]	0.2395 [0.1932, 0.2910]
Llama-3.1-8B	0.0553 [0.0329, 0.0808]	0.1844 [0.1496, 0.2218]	0.1292 [0.0970, 0.1658]
Mistral-7B-v0.3	0.0447 [0.0041, 0.0892]	0.2205 [0.1647, 0.2786]	0.1759 [0.1287, 0.2231]
Gemma-2-9B	0.0715 [0.0273, 0.1164]	0.2836 [0.2302, 0.3381]	0.2121 [0.1619, 0.2626]

Table 2: Cue-specific changes in population fit with 95% question-clustered bootstrap intervals. CI and EI are reductions in divergence from the same target distribution.

`question_id`, with random seed 42. Each sampled question retains all associated target populations, directed contrasts, reciprocal directions, cue conditions, and models, and the same cluster draw is shared across estimands. Question-level clustering preserves the dependence created by common item wording and matched observations. Row-wise resampling would incorrectly treat these records as independent. Conflict predictions are classified by the sign of $\mathrm{EO}_{\mathrm{raw}}$ using a numerical tolerance of 10^{-12}.

4 Results

We address three questions: how strongly labels and matched evidence improve population fit, which cue governs predictions under conflict, and how consistently the resulting hierarchy appears across models.

4.1 Country Labels Help, but Population Evidence Helps More

Figure 2 and Table 2 reveal a consistent separation between the two steering signals. Country labels produce positive mean Country Influence for all four models, ranging from 0.028 for Qwen to 0.071 for Gemma. The bootstrap interval excludes zero for Llama, Mistral, and Gemma, but not for Qwen. Country names therefore carry usable population-relevant information, although the effect is modest and not equally reliable across models.

Direct population evidence produces a substantially larger shift. Evidence Influence ranges from 0.184 to 0.284, with every bootstrap interval above zero. The paired evidence advantage, EI − CI, is also positive for every model and ranges from 0.129 for Llama to 0.240 for Qwen. Because CI and EI are computed against the same baseline and target distribution for each (q, L) unit, this comparison isolates the relative influence of the two cues rather than differences between survey questions or populations.

All four mean CI estimates are positive, indicating that country labels generally improve population fit. Direct population evidence produces a substantially larger shift, yielding a positive evidence advantage for every model. The model profiles differ within this shared ordering. Qwen has the smallest CI and the only interval spanning zero,

Model	$\mathrm{EO}_{\mathrm{raw}}$	$\mathrm{EO}_{\mathrm{norm}}$	Evidence-side / Label-side
Qwen2.5-7B	0.6384 [0.5714, 0.6954]	0.6686 [0.5916, 0.7394]	189 (94.5%) / 11 (5.5%)
Llama-3.1-8B	0.6147 [0.5563, 0.6631]	0.7114 [0.6416, 0.7712]	193 (96.5%) / 7 (3.5%)
Mistral-7B-v0.3	0.5167 [0.4273, 0.6018]	0.5433 [0.4408, 0.6421]	171 (85.5%) / 29 (14.5%)
Gemma-2-9B	0.6847 [0.6288, 0.7272]	0.7280 [0.6565, 0.7882]	195 (97.5%) / 5 (2.5%)

Table 3: Conflict arbitration over 200 directed contrasts per model. Intervals are clustered by question. Positive Evidence Override indicates that the prediction is closer to the evidence population. No prediction is tied at the prespecified tolerance.

Selection rule and abbreviated item	Label L	Evidence E	Qwen	Llama	Mistral	Gemma
Largest mean EO: sympathy for Muslims offended by cartoons	Spain	Jordan	0.940	0.976	0.940	0.941
Lowest mean EO: belief in God as necessary for morality	Sweden	Pakistan	-0.885	-0.973	-0.890	0.031
Largest model disagreement: Russia respects personal freedoms	Netherlands	Vietnam	-0.884	0.929	-0.899	0.909

Table 4: Mechanically selected conflict cases based on cross-model $\mathrm{EO}_{\mathrm{norm}}$. Positive values are evidence-side. Item descriptions are abbreviated, with complete prompts and distributions provided in the supplementary artifacts.

yet the largest paired evidence advantage, whereas Gemma has the largest mean CI and EI. Agreement on the hierarchy therefore does not imply equal sensitivity to either cue.

The same ordering holds within each survey source. The mean evidence advantage remains positive for every model when Global Attitudes and WVS contrasts are analyzed separately. This establishes robustness of the direction across the two source families without implying identical effect magnitudes.

4.2 When Names and Numbers Disagree, Evidence Usually Wins

The conflict condition directly tests cue arbitration by pairing country label L with evidence distribution p_E from a different population. All four models have positive mean raw and normalized Evidence Override, and every bootstrap interval lies above zero (Table 3). Mean normalized override ranges from 0.543 for Mistral to 0.728 for Gemma. The models therefore move substantially closer to the evidence population than to the named population.

Normalized override separates cue direction from update completeness. The sign records which reference is closer, whereas the magnitude records the strength of that relative proximity after normalizing for the distance between the references. A prediction can therefore be evidence-side while remaining far from p_E. This distinction explains why near-unanimous evidence-side classifications can coexist with mean $\mathrm{EO}_{\mathrm{norm}}$ values well below $+1$. Explicit evidence usually wins, but the model does not simply reproduce the displayed distribution.

The item-level classification reinforces the aggregate result. Evidence-side predictions account for 94.5% of Qwen conflicts, 96.5% of Llama conflicts, and 97.5% of Gemma conflicts. Mistral is less consistently evidence-dominant at 85.5%, but still follows the same overall hierarchy. None of the four models produces a tie at the prespecified tolerance.

4.3 The Cue Hierarchy Is Shared, but Not Uniform

The shared aggregate pattern is supported by substantial contrast-level agreement: all four models are evidence-side on 161 of the 200 directed contrasts, corresponding to 80.5% of the evaluation set. At least three models are evidence-side on 190 contrasts, or 95.0%. No contrast is label-side for all four models. The 161 all-model agreements therefore rule out an aggregation artifact in which each model follows evidence on a different subset of contrasts.

The agreement is nevertheless incomplete. Thirty-nine contrasts produce some cross-model disagreement. Twenty-nine of these are three-to-one splits, and Mistral is the sole label-side model in 20 cases. These disagreements localize a smaller region in which cue weighting is model-specific, consistent with Mistral's lower mean Evidence Override and larger label-side minority. The audit

thus supports a shared behavioral hierarchy without implying identical population representations.

Table 4 illustrates this variation using three deterministic selection rules. The contrast with the largest cross-model mean produces near-unanimous scores close to $+1$. The contrast with the lowest mean produces strong label-side predictions for Qwen, Llama, and Mistral, while Gemma lies just on the evidence side. The contrast with the greatest cross-model variance divides the models two against two.

The two difficult cases expose different structures. In the lowest-mean case, three models remain strongly label-side on the same contrast, showing that country anchoring can recur across model families. In the highest-variance case, the two-to-two split instead indicates model-specific arbitration. These cases complement the aggregate Evidence Override estimates by showing how model- and item-level heterogeneity coexists with a shared average direction.

5 Discussion

The results reveal a shared but incomplete hierarchy: country labels affect model predictions, but explicit population distributions govern conflict.

5.1 Country Labels as Defeasible Population Priors

The modest positive effects of country labels show that they carry usable population-level associations, while the larger evidence effects and conflict results place them lower in the cue hierarchy.

A parsimonious behavioral interpretation is that a country label acts as a weak, defeasible prior and the displayed distribution as stronger, question-specific evidence. This account characterizes observable cue weighting at the behavioral level and remains agnostic about the underlying computational mechanism. Mistral's larger proportion of label-side predictions further indicates that relative cue weighting is a model-specific behavioral property that warrants direct evaluation as a dimension distinct from general instruction-following performance.

5.2 From Steerability to Population Grounding

The audit connects steerable and distributional pluralism. Country-conditioned prompting demonstrates steerability when it changes behavior toward a named population, but population grounding requires that the change track the observed distribution of views within that population. Separating the label from the distribution turns this distinction into an empirical test: CI captures label responsiveness, EI captures matched evidence use, and conflict reveals which signal governs behavior when they disagree.

The evidence-dominant result is encouraging for pluralistic alignment because all four models translate an explicit population distribution into a corresponding shift over the same semantic options and usually prioritize it over a conflicting label. It also shows why country prompting cannot substitute for population data. A country name compresses heterogeneous social information and may continue to anchor predictions away from the observed sample.

The same distinction applies to age groups, professions, political constituencies, religions, and other internally diverse communities. In each case, evaluation should ask whether a model responds to the group label, and whether its response tracks evidence about the people denoted by that label.

5.3 From Direct Evidence Compliance to Generalization

The population-evidence condition reveals the target distribution for the same survey item. It therefore tests whether a model can use numerical population evidence and arbitrate it against a conflicting label. The strong evidence effects should be read as evidence compliance, not as proof of a transferable population model.

A harder extension would provide responses to related questions and evaluate held-out items while varying sample size, uncertainty, recency, and subgroup composition.

6 Conclusion

We have introduced a counterfactual audit that separates country labels from direct population evidence. Across open-weight models and directed contrasts, labels yield modest gains, evidence yields substantially larger gains, and evidence usually prevails under conflict. This pattern is shared across models, with model- and item-specific variation in update strength. Country names are informative but weaker cues. Evaluating pluralistic steering requires asking whether a model changes across populations and what information those changes track.

Limitations

We evaluate English prompts and four quantized open-weight models in the 7B–9B range. The results may not generalize to larger, closed, multilingual, or differently aligned systems. Quantization and GGUF conversion may also produce behavior that differs from the original publisher checkpoints.

GlobalOpinionQA distributions inherit survey sampling error, design choices, temporal variation, and uneven population coverage. National and territorial aggregates do not capture within-population heterogeneity and should not be treated as complete accounts of culture. Our starting set also combines value judgments with subjective political attitudes rather than following a single value taxonomy.

The selected contrasts maximize between-population divergence, making them useful for diagnosing cue arbitration but unrepresentative of the full dataset. The direct evidence condition displays the target distribution and therefore tests evidence use rather than held-out population inference. We use one fixed option permutation for matched comparison rather than averaging over multiple orders. Finally, this is a behavioral audit. Prompt interventions identify which cue governs the observed prediction, but not the model's internal representation or computational mechanism.

Ethical Considerations

The study uses publicly released aggregate survey distributions and processes no personally identifying information. Nevertheless, country-conditioned modeling can encourage cultural essentialism, stereotyping, or demographic profiling. We use country and territory names only as controlled prompt cues, and we treat survey distributions as sample-dependent population references rather than definitions of culture or normative truths.

The conflict condition deliberately pairs one population label with another population's observed distribution. These prompts are counterfactual diagnostic interventions and must not be presented as factual descriptions of either population. Agreement with a population aggregate also provides no basis for inferring the values of an individual from nationality or residence.

To support reproducibility, we publicly release the analysis code, processed data, prompts, exclusion records, provenance information and validated results at `https://github.com/CoconutKerry/country-name-or-cultural-evidence`. The analysis code is released under the MIT License. The processed data derive from GlobalOpinionQA (DURMUS et al., 2024), which draws on the Pew Global Attitudes Survey and the World Values Survey, and are provided for research use only, subject to the applicable source-data terms. We do not redistribute model weights. The evaluated models and the llama.cpp inference framework remain subject to their respective licenses.

Acknowledgments

The present research was supported by Industrial Key Core Technology Research in the Guangdong-Macao In-Depth Cooperation Zone in Hengqin (Project No. 2530004025501). We would like to thank the anonymous reviewers for their insightful comments.

References

M. Farid Adilazuarda, Chen Cecilia Liu, Iryna Gurevych, and Alham Fikri Aji. 2025. From surveys to narratives: Rethinking cultural value adaptation in LLMs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 18052–18079, Suzhou, China. Association for Computational Linguistics.

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 (Volume 1: Long Papers)*, pages 12404–12422, Bangkok, Thailand. 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.

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, Albuquerque, New Mexico. Association for Computational Linguistics.

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*, volume 37, pages 45850–45878. Curran Associates, Inc.

Esin DURMUS, Karina Nguyen, Thomas Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, Liane Lovitt, Sam McCandlish, Orowa Sikder, Alex Tamkin, Janel Thamkul, Jared Kaplan, Jack Clark, and Deep Ganguli. 2024. Towards measuring the representation of subjective global opinions in language models. In *First Conference on Language Modeling*.

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

Gemma Team. 2024. Gemma 2: Improving open language models at a practical size. *Preprint*, arXiv:2408.00118.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The Llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7B. *Preprint*, arXiv:2310.06825.

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*, pages 24–51, Suzhou, China. Association for Computational Linguistics.

Ariba Khan, Stephen Casper, and Dylan Hadfield-Menell. 2025. Randomness, not representation: The unreliability of evaluating cultural alignment in llms. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '25, page 2151–2165, New York, NY, USA. Association for Computing Machinery.

Jindřich Libovický. 2026. On the credibility of evaluating LLMs using survey questions. In *Proceedings of the First Workshop on Multilingual Multicultural Evaluation*, pages 23–34, Rabat, Morocco. Association for Computational Linguistics.

Yifan Lyu, Xinran Li, Jiaqi Qiao, and Xiujuan Xu. 2026. Signal or spurious cue? a randomized audit of survey-country metadata in llm social inference. *Preprint*, arXiv:2608.06085.

Viet Thanh Pham, Lizhen Qu, Zhuang Li, and Gholamreza Haffari. 2026. Distributional alignment for large language models under domain shift. In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 20514–20528, San Diego, California, United States. Association for Computational Linguistics.

Qwen Team. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Paul Röttger, Valentin Hofmann, Valentina Pyatkin, Musashi Hinck, Hannah Kirk, Hinrich Schuetze, and Dirk Hovy. 2024. Political compass or spinning arrow? towards more meaningful evaluations for values and opinions in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15295–15311, Bangkok, Thailand. Association for Computational Linguistics.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose opinions do language models reflect? In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 29971–30004. PMLR.

Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell L Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, Tim Althoff, and Yejin Choi. 2024. Position: A roadmap to pluralistic alignment. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 46280–46302. PMLR.

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

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

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

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.

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

Pengyun Zhu, Yuqi Ren, Zhen Wang, Lei Yang, and Deyi Xiong. 2026. DVMap: Fine-grained pluralistic value alignment via high-consensus demographic-value mapping. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 19834–19852, San Diego, California, United States. Association for Computational Linguistics.